THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXV NOVEMBER 1956 NUMBER 6

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at $3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U.S.A.

Nobel Prize in Physics Awarded to Transistor Inventors

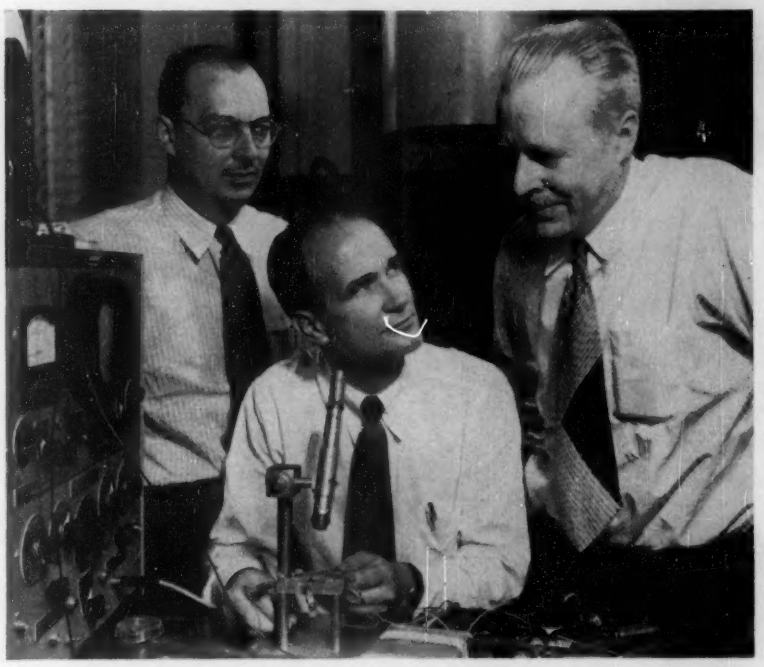
The Swedish Royal Academy of Sciences announced on November 1 that a Nobel Prize in Physics, most highly coveted award in the world of physics, had been awarded jointly to Dr. Walter H. Brattain of the Laboratories Physical Research Department, with Dr. John Bardeen and Dr. William Shockley, both former members of the Laboratories. The prize was awarded for "investigations on semiconductors and the discovery of the transistor effect."

This marks the second time that work done at the Laboratories has been recognized by a Nobel Prize. The previous recipient was Dr. C. J. Davisson who shared in the 1937 prize for his discovery of electron diffraction as a result of experiments carried out with Dr. L. H. Germer, also of the Laboratories.

Each of the three winners of this year's prize will receive a gold medal, a diploma and a share of the $38,633 prize money. When he was notified that he was one of these winners, Dr. Brattain said, "I certainly appreciate the honor. It is a great satisfaction to have done something in life and to have been recognized for it in this way. However, much of my good fortune comes from being in the right place, at the right time, and having the right sort of people to work with."

The principle of transistor action was discovered as a result of fundamental research directed toward gaining a better understanding of the surface properties of semiconductors. Following World War II, intensive programs on the properties of germanium and silicon were undertaken at the Laboratories under the direction of William Shockley and S. O. Morgan. One group in this program engaged in a study of the body properties of semi-conductors, and another on the surface properties. Dr. John Bardeen served as theoretical physicist and R. B. Gibney as chemist for both groups. These investigations, which resulted in the invention of the transistor, made extensive use of knowledge and techniques developed by scientists here and elsewhere, particularly by members of the Laboratories—R. S. Ohl, J. H. Scaff and H. C. Theuerer.

Since the transistor was announced, little more than eight years ago, it has become increasingly important in what has been called the "new



The Nobel Prize winners in an historic photograph taken in 1948 when the announcement of the invention of the transistor was made. Left to right, John Bardeen, William Shockley and Walter H. Brattain.

electronics age." As new transistors and related semiconductor devices are developed and improved, the possible fields of application for these devices increase to such an extent that they may truly be said to have "revolutionized the electronics art."

The invention of the transistor, basis for the Nobel Prize award, represents an outstanding example of the combination of research teamwork and individual achievement in the Bell System that has meant so much to the rapid development of modern communications systems.

Dr. Brattain received a B.S. degree from Whitman College in 1924, an M.A. degree from the University of Oregon in 1926, and a Ph.D. degree from the University of Minnesota in 1928. He joined Bell Telephone Laboratories in 1929, and his early work was in the field of thermionics, particularly the study of electron emission from hot surfaces. He also studied frequency standards, magnetometers and infra-red phenomena.

Subsequently, Mr. Brattain engaged in the study of electrical conductivity and rectification phenomena in semiconductors. During World War II, he was associated with the National Defense Research Committee at Columbia University where he worked on magnetic detection of submarines.

Mr. Brattain has received honorary Doctor of Science degrees from Whitman College, Union College and Portland University. His many awards include the John Scott Medal and the Stuart Ballantine Medal, both of which he received jointly with John Bardeen. Mr. Brattain is a Fellow of the American Academy of Arts and Sciences.

Dr. Bardeen received the B.S. in E.E. and M.S. in E.E. degrees from the University of Wisconsin in 1928 and 1929 respectively, and his Ph.D. degree in Mathematics and Physics from Princeton University in 1936. After serving as an Assistant Professor of Physics at the University of Minnesota from 1938 to 1941, he worked with the Naval Ordnance Laboratory as a physicist during World War II. In 1945 he joined the Laboratories as a research physicist, and was primarily concerned

Clinton J. Davisson Previous Laboratories Nobel Laureate

In December, 1937, Dr. Clinton J. Davisson of the Laboratories was awarded the Nobel Prize in Physics for his discovery of electron diffraction and the wave properties of electrons.

He shared the prize with Professor G. P. Thompson of London, who worked in the same field, though there was little in common between their techniques. Dr. Davisson's work on electron diffraction started as an attempt to understand the characteristics of secondary emission in multigrid electron tubes. In this work he discovered patterns of emission from the surface of single crystals of nickel. By studying these patterns, Dr. Davisson, with Dr. L. H. Germer and their associates, proved that reflected electrons have the properties of trains of waves.

Dr. Davisson was awarded the B.S. degree in physics from the University of Chicago in 1908 and the Ph.D. degree from Princeton in 1911. From September, 1911, until June, 1917, he was an instructor in physics at the Carnegie Institute of Technology, coming to the Laboratories on a wartime leave of absence. He found the climate of the Laboratories conducive to basic research, however, and remained until his retirement in 1946. Besides his work on electron diffraction, Dr. Davisson did much significant work in a variety of fields, particularly electron optics, magnetrons, and crystal physics.

with theoretical problems in solid state physics, including studies of semiconductor materials.

Mr. Bardeen, whose honors include an honorary Doctor of Science degree from Union College, the Stuart Ballantine Medal, the John Scott Medal, and the Buckley Prize, is a member of the National Academy of Sciences. He joined the University of Illinois in 1951.

Dr. Shockley received a B.Sc. degree from the California Institute of Technology in 1932, and a Ph.D. degree from the Massachusetts Institute of Technology in 1936. He joined the staff of Bell Telephone Laboratories in 1936. In addition to his many contributions to solid state physics and semiconductors, Mr. Shockley has worked on electron tube and electron multiplier design, studies of various physical phenomena in alloys, radar development and magnetism.

His many awards include an honorary degree from the University of Pennsylvania, the Morris Liebmann Memorial Prize, the Buckley Prize, the Comstock Prize and membership in the National Academy of Sciences. Dr. Shockley left the Laboratories to form the Shockley Semiconductor Laboratory at Beckman Instruments, Inc., in 1955.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXV NOVEMBER 1956 NUMBER 6

Theory of the Swept Intrinsic Structure

By. W. T. READ, JR.

(Manuscript received March 4, 1956)

The electric field and the hole and electron concentrations are found for reverse biased junctions in which one side is either intrinsic (I) or so weakly doped that the space charge of the carriers cannot be neglected. The analysis takes account of space charge, drift, diffusion and non linear recombination. A number of figures illustrate the penetration of the electric field into a PIN structure with increasing bias for various lengths of the I region. For the junction between a highly doped and a weakly doped region, the reverse current increases as the square root of the voltage at high voltages; and the space charge in the weakly doped region approaches a constant value that depends on the fixed charge and the intrinsic carrier concentration.

The mathematics is greatly simplified by expressing the equations in terms of the electric field and the sum of the hole and electron densities.

I. INTRODUCTION

Applications have been suggested for semiconductor structures having both extrinsic and intrinsic regions. Examples are the "swept intrinsic" structure, in which a region of high resistivity is set up by an electric field that sweeps out the mobile carriers, and the analogue transistors, where the intrinsic region is analogous to the vacuum in a vacuum tube. However, the junction between an intrinsic region and an N or P region

is considerably less well understood than the simple NP junction. Most of the assumptions that make the NP case relatively simple to deal with do not apply to junctions where one side is intrinsic. Specifically, the space charge is that of the mobile carriers; thus the flow and electrostatic problems cannot be separated as they can in PN junction under reverse bias. The following sections analyze the N-intrinsic - P structure under reverse bias.

For a given material with fairly highly doped extrinsic regions, the problem is defined by the length of the intrinsic region and the applied voltage. Taking the intrinsic region infinitely long gives the solution for a simple N-intrinsic or P-intrinsic structure. The results are given and plotted in terms of the electric field distribution. From this the potential, space charge and carrier concentrations can be found; so also can the current-voltage curve. The final section considers the case where the middle layer contains some fixed charge but where the carrier charge cannot be neglected.

Qualitative Discussion of an N-intrinsic-P Structure

Consider an N-intrinsic-P structure where the intrinsic, or I, region is considerably wider than the zero bias, or built-in, space charge regions at the junctions, so that there is normal intrinsic material between the junctions. The field distribution at zero bias can be found exactly from the zero-current analysis of Prim.[1] Throughout the intrinsic region, hole and electron pairs are always being thermally generated and recombining at a rate determined by the density and properties of the traps, or recombination centers. Under zero bias the rates of generation and recombination are everywhere equal. Suppose now a reverse bias is applied causing holes to flow to the right and electrons to the left. Some of the carriers generated in the intrinsic region will be swept out before recombining. This depletes the carrier concentration in the intrinsic region and hence raises the resistivity. It also produces a space charge extending into the intrinsic layer. The electrons are displaced to the left and the holes, to the right. Thus the space charge opposes the penetration of the field into the intrinsic region; that is, the negative charge of the electrons on the left and positive charge of the holes on the right gives a field distribution with a minimum somewhere in the interior of the intrinsic region and maxima at the NI and IP junctions. If holes and electrons had equal mobilities, the field distribution would be symmetrical with a minimum in the center of the intrinsic region. Likewise, the total carrier

[1] R. C. Prim, B. S. T. J., **32,** p. 665, May, 1953.

concentration (holes plus electrons) would be symmetrical with a maximum in the center. As the applied bias is increased the hole and electron distributions are further displaced relative to one another and the space charge increases. Finally, at high enough biases, so many of the carriers are swept out immediately after being generated that few carriers are left in the intrinsic region. Now the space charge decreases with increasing bias until there is negligible space charge, and a relatively large and constant electric field extends through the intrinsic region from junction to junction. This may happen at biases that are still much too low to appreciably affect the high fields right at the junction or in the extrinsic layers, which remain approximately as they were for zero bias.

The current will increase with voltage until the total number of carriers in the intrinsic region becomes small compared to its normal value. After that, there is negligible further increase of current with voltage. All the carriers generated in the intrinsic region are being swept out before recombining. In general, the current will saturate while the minimum field in the intrinsic region is still small compared to the average field.

Comparison with the NP Structure

The analysis is more difficult than in a simple reverse-biased *NP* structure. In the *NP* case there is a well defined space charge region in which carrier concentration is negligible compared to the fixed charge of the chemical impurities; so the field and potential distributions are easily found from the known distribution of fixed charge. Outside of the space charge region are the diffusion regions in which the minority carrier concentration rises from a low value at the edge of the space charge region to its normal value deep in the extrinsic region. However, there is no space charge in this region because the majority carrier concentration, by a very small percentage variation, can compensate for the large percentage variation in minority carrier density. The minority carriers flow by diffusion. Since the disturbance in carrier density is small compared to the majority density, the recombination follows a simple linear law (being directly proportional to the excess of minority carriers). Thus the minority carrier distribution is found by solving the simple diffusion equation with linear recombination.

None of these simplifications extend to the *NIP* or *NI* or *IP* structure. There is, in the intrinsic region, no fixed charge; hence the space charge is that of the carriers. There is no majority carrier concentration to maintain electrical neutrality outside of a limited space charge region.

It is necessary to take account of (1) space charge, (2) carrier drift, (3) carrier diffusion and (4) recombination according to a nonlinear bimolecular law. Of these four, only space charge and recombination are never simultaneously important in practical cases. Nevertheless certain simplifications can be made if the problem is formulated so as to take advantage of them. The field and carrier distributions in the intrinsic region are found by joining two solutions: one solution is for charge neutrality; the other, which we shall call the no-recombination solution is for the case where the recombination rate is negligible compared to the rate of thermal generation of hole electron pairs. We shall show that in practical cases the ranges of validity of the two solutions overlap; that is, wherever recombination is important, we have charge neutrality.

Prim's Zero-Current Approximation

Prim* derived the field distribution in a reverse biased *NIP* structure on the assumption that the hole and electron currents are negligibly small differences between their drift and diffusion terms, as in the zero-bias case. He showed that the average diffusion current is large compared to the average current. However, as it turns out, this is misleading. Throughout almost all of the intrinsic region (where the voltage drop occurs in practical cases) the diffusion current is comparable to or smaller than the total current. The larger average diffusion current comes from the extremely large diffusion current in the small regions of high space charge at the junctions. Prim's analysis, in effect, neglects the space charge of the carriers generated in the intrinsic region. These may be neglected in calculating the field distribution if the intrinsic region is sufficiently narrow or the reverse bias sufficiently high. In the appendix we derive the limits within which Prim's calculation of the field and potential will be valid. The range will increase with both the Debye length and the diffusion length in the intrinsic material. However, in cases of practical interest the zero-current approximation may lead to serious errors in the field distribution and give a misleading idea of the penetration of the field into the intrinsic region. The present, more general analysis, reduces to Prim's near the junctions where the zero-current assumption remains valid. The zero current approximation was, of course, not intended to give the hole and electron distributions in the intrinsic region or to evaluate the effects of interacting drift, diffusion and recombination.

---

* Ibid.

Outline of the Following Sections

Sections II through V deal with the ideal case of equal hole and electron mobilities. Here the problem is somewhat simplified and the physics easier to visualize because of the resulting symmetry. In Section VI, the general case of arbitrary mobilities is solved by an extension of the methods developed for solving the ideal case. The technique is to deal not with the hole and electron flow densities but with two linear combinations of hole and electron flow densities that have a simple form.

Section II deals with the basic relations and in particular the formula for recombination in an intrinsic region for large disturbances in carrier density. The nature and range of validity of the various approximations are discussed. Section III derives the field distribution in regions where recombination is small compared to pair generation. Section IV treats the recombination region and the smooth joining of the recombination and no-recombination solutions. Section V considers the role of diffusion in current flow and the situation at the junctions where the field and carrier concentration abruptly become large. The change in form of the solution near the junctions is shown to be represented by a basic instability in the governing differential equation. Section VI extends the results to the general case of unequal mobilities. Section VII deals with the still more general case where there is some fixed charge in the "intrinsic" region. If the density of excess chemical impurities is small compared to the intrinsic carrier density, the solution remains unchanged in the range where recombination is important. In the no-recombination region the solution is given by a simple first order differentiatial equation which can be solved in closed form in the range where the carrier flow is by drift. The fixed charge may have a dominant effect on the space charge even when the excess density of chemical impurities is small compared to the density n_i of electrons in intrinsic material. Consider, for example, a junction between an extrinsic P region and a weakly doped n region having an excess density $N = N_d - N_a$ of donors. In the limit, as the reverse bias is increased and the space charge penetrates many diffusion lengths into the n region, the field distribution becomes linear, corresponding to a constant charge density equal to

$$\tfrac{1}{2}[N + \sqrt{N^2 + 8\, n_i^2 \mathfrak{L}^2/L_i^2}]$$

where L_i is the diffusion length in the weakly doped n type region and $\mathfrak{L}$ is the Debye length for intrinsic material. For germanium at room temperature $\mathfrak{L}/L_i$ is the order of 10^{-3}. Thus, in this example, a donor density as low as 10^{11} cm^{-3} will have an appreciable effect on the space charge.

II. BASIC RELATIONS

The problem can be stated in terms of the hole density p, the electron density n, and the electric field E and their derivatives. Let the distance x be measured in the direction from N to P. The field will be taken as positive when a hole tends to drift in the $+x$ direction. The field increases in going in the $+x$ direction when the space charge is positive. Poisson's equation for intrinsic material is

$$\frac{dE}{dx} = a(p - n) \tag{2.1}$$

where the constant a has the dimensions of volt cm and is given in terms of the electronic charge q and the dielectric constant κ by

$$a = \frac{4\pi q}{\kappa}$$

For germanium $a = 1.17 \times 10^{-7}$ volt cm.

The hole and electron flow densities J_p and J_n are[2]

$$\begin{aligned} J_p &= \mu E p - D\frac{dp}{dx} = \mu p\left[E - \frac{kT}{q}\frac{d}{dx}\ell n\, p\right] \\ J_n &= -b\left(\mu E n + D\frac{dn}{dx}\right) = -b\mu n\left[E + \frac{kT}{q}\frac{d}{dx}\ell n\, n\right] \end{aligned} \tag{2.2}$$

where μ and $D = \mu\, kT/q$ are the hole mobility and diffusion constant respectively, k is Boltzmann's constant (8.63×10^{-5} ev per °C) and T is the absolute temperature. The ratio b of electron mobility to hole mobility we take to be unity. This makes the problem symmetrical in n and p and consequently easier to understand. Section VI will extend the results to the general case of arbitrary b.

Charge and Particle Flow

For some purposes it helps to express the flow not in terms of J_p and J_n but rather in terms of the current density I and the flow density $J = J_p + J_n$ of particles, or carriers. The current density $I = q(J_p - J_n)$. Each carrier, hole or electron, gives a positive contribution to J if it goes in the $+x$ direction and a negative contribution if it goes in the $-x$ direction. In other words, J is the net flow of carriers regardless of their charge sign. The current I is constant throughout the intrinsic

[2] See, for example, Electrons and Holes in Semiconductors, by W. Shockley. D. Van Nostrand Co., New York, 1950.

region. Particle flow is away from the center of the intrinsic region. Carriers are generated in the intrinsic region and flow out at the two ends, the electrons going out on the N side and holes on the P side. Thus J is positive near the IP junction and negative near the NI junction.

From the definitions of I and J and equations (2.2)

$$\begin{aligned} \frac{I}{q} &= \mu E(p + n) - D\frac{d}{dx}(p - n) \\ J &= \mu E(p - n) - D\frac{d}{dx}(p + n) \end{aligned} \tag{2.3}$$

It is convenient to express the equations in terms of E and a dimensionless variable

$$s = \frac{n + p}{2n_i} \tag{2.4}$$

which measures how "swept" the region is. In normal intrinsic material $s = 1$. In a completely swept region $s = 0$; at the junctions with highly extrinsic material $s \gg 1$. Using Poisson's equation to express $p - n$ in terms of E, equations (2.3) become

$$\begin{aligned} I &= \sigma s E - \frac{qD}{a}\frac{d^2E}{dx^2} \\ J &= \frac{d}{dx}\left[\frac{\mu E^2}{2a} - 2n_i D s\right] \end{aligned} \tag{2.5}$$

where $\sigma = 2\,\mu\, n_i q$ is the conductivity of intrinsic material. The particle flow J is thus seen to be the gradient of a flow potential that depends only on E and s.

Equations (2.5) can be written in the form

$$I = \sigma\left[sE - \mathfrak{L}^2\frac{d^2E}{dx^2}\right] \tag{2.6}$$

$$J = D2n_i\frac{d}{dx}\left[\frac{E^2}{E_1^2} - s\right] \tag{2.7}$$

where $\mathfrak{L} = \sqrt{kT/2an_iq}$ is the Debye length in intrinsic material and

$$E_1 = 2\sqrt{\frac{an_ikT}{q}} = \frac{\sqrt{2}kT}{q\mathfrak{L}} \tag{2.8}$$

is a field characteristic of the material and temperature. Specifically E_1 is $\sqrt{2}$ times the field required to give a voltage drop kT/q in a Debye

length. For germanium at room temperature $\mathfrak{L} = 6.8\ 10^{-5}$ cm and $E_1 = 383$ volts per cm.

Both I and J are the sum of a drift term and a diffusion term. For charge neutrality, where $p - n$ is small compared to $p + n$, both charge diffusion and particle drift can be neglected. We shall see later that, except right at the junctions, charge diffusion is negligible.

The Equations of Continuity

The two equations of continuity are

$$\frac{dJ_p}{dx} = \frac{dJ_n}{dx} = g - r \tag{2.9}$$

where g is the rate of pair generation and r the rate of recombination. In terms of I and J, these become

$$\frac{dI}{dx} = 0 \tag{2.10}$$

or $I =$ constant and

$$\frac{dJ}{dx} = 2(g - r) \tag{2.11}$$

which says that the gradient of particle flow is equal to the net rate of particle generation, that is, twice the net rate of pair generation.

To complete the statement of the problem it remains to express g and r in terms of n and p.

Generation and Recombination

The direct generation and recombination of holes and electrons follows the mass action law, in which $g - r$ is proportional to $n_i^2 - np$. The constant of proportionality can be defined in terms of a lifetime τ as follows: Let $\delta p = \delta n \ll n_i$ be a small disturbance in carrier density. Then defining $\tau(g - r) = -\delta n$, we see that the proportionality constant in the mass action law is $(2n_i\tau)^{-1}$. So

$$g - r = \frac{n_i^2 - np}{2n_i\tau} \tag{2.12}$$

and the generation rate

$$g = \frac{n_i}{2\tau} \tag{2.13}$$

is independent of carrier concentration.

In actual semiconducting materials, recombination is not direct. Rather it occurs through a trap, or recombination center. The statistics of indirect recombination has been treated by Shockley and Read[3] for a recombination center having an arbitrary energy level $\mathcal{E}_t$ somewhere in the energy gap. At any temperature the trap level can be expressed by the values n_1 and p_1 which n and p would have if, at that temperature, the Fermi level were at the trap level. Shockley and Read showed that, at a given temperature, the lifetime for small disturbances in carrier density is a maximum in intrinsic material. It drops to limiting values τ_{n0} and τ_{p0} in highly extrinsic n and p material, respectively. The formula for $g - r$ in terms of n and p is

$$g - r = \frac{n_i^2 - np}{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)} \tag{2.14}$$

For our purposes it is more convenient to define the lifetime τ not by $\tau(g - r) = -\delta n \ll n_i$, but rather as the proportionality factor in the mass action law. Then τ is not necessarily constant independent of carrier density. From (2.12) and (2.14)

$$\tau = \frac{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)}{2n_i} \tag{2.15}$$

We shall be interested in the lifetime in the region where n and p are equal to or less than n_i. τ decreases as n and p decrease; that is, τ is less in a swept region than in normal intrinsic material. Let $\tau = \tau_i$ for $n = p = n_i$ and $\tau = \tau_0$ for $n = p = 0$. The total range of variation of τ is by a factor of

$$\frac{\tau_i}{\tau_0} = 1 + \frac{n_i(\tau_{p0} + \tau_{n0})}{p_1\tau_{n0} + n_1\tau_{p0}} \tag{2.16}$$

Let the energy levels be measured relative to the intrinsic level, and define a level $\mathcal{E}_0$ by

$$\mathcal{E}_0 = kT \ln \sqrt{\frac{\tau_{n0}}{\tau_{p0}}}$$

Then if $\mathcal{E}_t = \mathcal{E}_0$, $n_1\tau_{p0} = p_1\tau_{n0}$. Now eq. (2.16) becomes

$$\frac{\tau_i}{\tau_0} = 1 + \frac{1}{2}\left(\sqrt{\frac{\tau_{n0}}{\tau_{p0}}} + \sqrt{\frac{\tau_{p0}}{\tau_{n0}}}\right) \operatorname{sech}\left(\frac{\mathcal{E}_t - \mathcal{E}_0}{kT}\right) \tag{2.17}$$

Thus the variation in τ increases as the ratio of τ_{n0} to τ_{p0} deviates from unity and as the trap level moves away from the level $\mathcal{E}_0$.

[3] W. Shockley and W. T. Read, Jr., Phys. Rev., **87**, p. 835, 1952.

The data of Burton, Hull, Morin, and Severien[4] shows that a typical value of the ratio of τ_{p0} and τ_{n0} is about 10. This means that the variation in τ with carrier concentration will be less than 10 per cent provided $\mathcal{E}_f$ is about $4kT$ from $\mathcal{E}_0$. In what follows we shall assume that this is so. Then we have the mass action law (2.12) with τ a constant, which could be measured by one of the standard techniques involving small disturbances in carrier density. The general case of variable τ is considered briefly at the end of Section IV.

Outline of the Solution

To conclude this section, we discuss briefly the form of the equations and the solution in different parts of the intrinsic region. First consider (2.6) for the current in the ideal case of equal mobilities. In Sections III and V we shall show that throughout almost all of the intrinsic region the current flows mainly by pure drift so we can take $I = \sigma s E$. The reason for this is as follows. The quantity $\mathfrak{L}^2$ is so small that the diffusion term remains negligible unless the second derivative of E becomes large — so large in fact that the E versus x curve bends sharply upward and both the drift and diffusion terms become large compared to the current I. This is the situation at the junction where I is the small difference between large drift and diffusion terms. Thus (2.6) has two limiting forms:

(1) Except at the junctions the current is almost pure drift so $I = \sigma s E$ is a good approximation. In Section III we derive an upper limit for the error introduced by this approximation and show how the approximate solution can be corrected to take account of the diffusion term.

(2) At the junction, the drift term becomes important and the current rapidly becomes a small difference between its drift and diffusion terms and the solution approaches the zero current solution, for which $sE = \mathfrak{L}^2 d^2E/dx^2$. In Section V we derive an approximate solution that joins onto the $I = \sigma s E$ solution near the junction and then turns continously and rapidly into the zero current solution. We shall call this the *junction* solution.

The abrupt change in the solution from (1) to (2) near the junction is shown to be related to a basic instability in the differential equation. This makes it impractical to solve the equations on a machine.

When the applied bias is large compared to the built-in voltage drop, the junction region will be of relatively little interest so the $I = \sigma s E$ solution can be used throughout.

In the region where $I = \sigma s E$ there are two overlapping regions in which the equations assume a simple form. These are the following:

[4] Burton, Hull, Morin and Severiens, J. Phys. Chem., **57**, p. 853, 1953.

The No-Recombination Solution

Here recombination is small compared to generation, $r \ll g$. This will be so in at least part of the intrinsic region for reverse biases of more than a few kT/q. The E versus x curve turns out to be given by a simple, cubic algebraic equation.

The Recombination, or Charge Neutrality, Solution

Here $n - p$ is small compared to $n + p$, so the particle flow is by diffusion. We shall find that the s versus x curve is given by a third degree elliptic integral. As we move away from the center of the intrinsic region and toward the junctions, recombination becomes small compared to generation and the recombination solution goes into the no-recombination solution. In the region where both solutions hold, the solution has the simple form $s = I/\sigma E = A - x^2$ where A is a constant that must be less than $\frac{3}{4}$ and the unit of length is twice the diffusion length.

As the bias on an NIP structure is increased and the space charge penetrates through the intrinsic region, the region where the recombination is important will shrink and eventually disappear.

Fig. 1 is a schematic plot of the field distribution for the case where the applied bias is large compared to the built-in potential drop but not large enough to sweep all the carriers out of the intrinsic region. As the voltage is increased, the drop in field in the intrinsic region will become less and finally the field distribution will be almost flat from junction to junction. Only half of the intrinsic region is shown in Fig. 1. For equal mobilities the field distribution will be symmetrical about the center x_i of the intrinsic region.

The illustration shows the recombination solution (1), which holds near the center of the intrinsic region and overlaps (2), the no-recombination solution. The junction solution (3) joins continuously onto the no-recombination solution at the point x_0 and rapidly breaks away and approaches the zero-current solution at the junction. The figure is schematic and has not been drawn to scale. In most cases of interest, the low fields in the recombination region will be much lower and the junction solution will hold over a smaller fraction of the intrinsic region.

It is convenient to take $x = 0$ not at the center x_i of the intrinsic region but at the minimum on the no-recombination solution. As the applied bias increases, x_i approaches zero.

Unequal Mobilities

In the general case of unequal mobilities, it is no longer so that I is pure drift except at the junctions. However we can define a linear com-

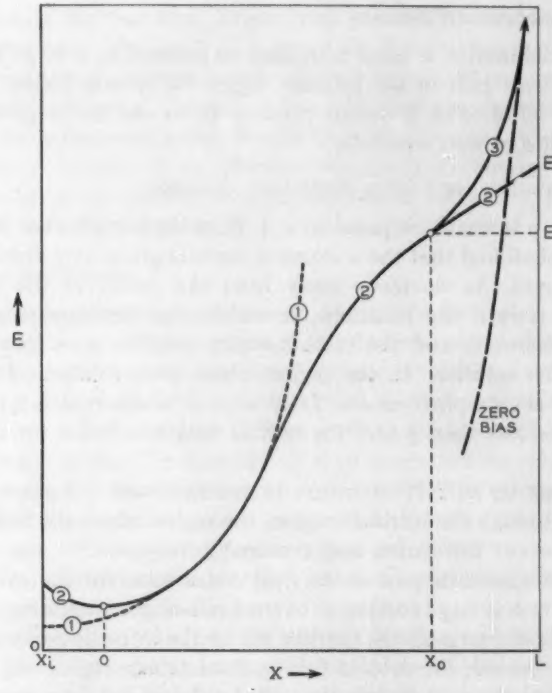


Fig. 1 — Schematic of the field distribution and the three overlapping solutions.

bination of J_p and J_n which has the same form as I in (2.6) and in which the diffusion term is negligible except near the junction. As we show in section VI, the effect of unequal mobilities is (1) to introduce some asymmetry into the curve in the region where the curvature is upward and (2) to displace the curve toward the NI junction (for the case where the electrons have the higher mobility). Thus the field is higher on the side where the carrier mobility is lower, as would be expected.

III. THE NO-RECOMBINATION CASE

This section deals with the case where recombination can be neglected in comparison with generation. This will be so where np is small compared to n_i^2.

The continuity equation for J now becomes

$$\frac{dJ}{dx} = 2g = \frac{n_i}{\tau} \tag{3.1}$$

Combining this with (2.7) gives

$$\frac{d^2}{dx^2}\left(\frac{E^2}{E_1^2} - s\right) = \frac{1}{2L_i^2} \tag{3.2}$$

where $L_i^2 = D\tau$ is the diffusion length in intrinsic material.

Equation (3.2) can be immediately integrated. There are two constants of integration, one of which can be made to vanish by choosing $x = 0$ at the center of the intrinsic region, where the first derivatives of E and s vanish. (E is a minimum here and s a maximum). The solution obtained by two integrations is

$$\left(\frac{E}{E_1}\right)^2 - s = \left(\frac{x}{2L_i}\right)^2 - A \tag{3.3}$$

As we shall see later, the constant A is determined by the voltage drop across the unit.

The exact procedure now would be to substitute s from (3.3) into (2.6). The resulting second order differential equation could, in principle, then be solved for E versus x. The exact solution, however, would be quite difficult. We shall discuss it in Section V. Here we make the assumption that throughout the intrinsic region the charge flow is mainly by drift, so that we can neglect the diffusion term in (2.6) and take $I = \sigma s E$, as discussed in Section II. Later in this section we find an upper limit on the error due to this assumption and show how the cubic can be corrected to take account of the diffusion term.

Putting $s = I/\sigma E$ in (3.3) gives a cubic equation

$$\begin{aligned}\left(\frac{E}{E_1}\right)^2 - \frac{I}{\sigma E} &= \left(\frac{x}{2L_i}\right)^2 - A \\ \left(\frac{E}{E_1}\right)^2 - \frac{I}{\sigma E_1}\left(\frac{E_1}{E}\right) &= \left(\frac{x}{2L_i}\right)^2 - A\end{aligned} \tag{3.4}$$

for E/E_1 as a function of $x/2L_i$. This equation contains two parameters I and A. A determines the voltage and I is determined by the length $2L$ of the intrinsic region. The relation is as follows: Let the applied voltage drop across each junction be at least a few kT/q. Then the minority carrier currents from the extrinsic regions will have reached their saturation values. Call I_s the sum of the hole current from the N region and the electron current from the P region. I_s comes from pairs generated in the extrinsic regions near the junctions. I_s can be made arbitrarily small by making the N and P regions sufficiently highly doped (provided the diffusion length in the extrinsic material does not decrease with doping faster than the majority carrier concentration in-

creases). The current generated in the intrinsic region is qg per unit volume. So the density of current from pairs generated in the intrinsic layer is $2Lqg = qn_iL/\tau$. Hence

$$I = I_s + \frac{qn_iL}{\tau}$$

In what follows we shall assume that I_s is negligibly small compared to I. Then

$$I = \left(\frac{qn_i}{\tau}\right) L = \left(\frac{qn_iD}{L_i}\right) \frac{L}{L_i} = \left(\frac{\sigma}{2L_i} \frac{kT}{q}\right) \frac{L}{L_i}$$

Thus I is L/L_i times a characteristic current equal to (1) the diffusion current produced by a gradient n_i/L_i or (2) the drift current produced by a field that gives the voltage drop kT/q in two diffusion lengths in normal intrinsic material. In germanium this characteristic current is about 5 milliamperes per cm^2.

That the current I is proportional to L and independent of voltage follows from the neglect of recombination. When recombination is small compared to generation, then the current has reached its maximum, or saturation, value. All the carriers generated in the intrinsic region are swept out before recombining. It will sometimes be convenient to take σE_1 as the unit of current. From the above and (2.8)

$$\frac{I}{\sigma E_1} = \frac{\sqrt{2}\mathfrak{L}L}{(2L_i)^2} \tag{3.5}$$

In germanium σE_1 is about 7 amperes per cm^2. In general we will be dealing with currents that are small compared to this. For example, if L_i is 1 mm, we would have to sweep out an intrinsic region 3 meters long in order to get a current this large. If we take E_1 as the unit field, σE_1 as the unit current and $2L_i$ as the unit length then the cubic becomes $E^2 - I/E = x^2 - A$.

For a given structure and temperature the field versus x curves form a one parameter family. A determines both the field distribution and the voltage. The voltage increases as A decreases. Fig. 2 is a plot of E/E_1 versus $x/2L_i$ for $L/2L_i = 0.1$ and several different values of A. Fig. 3 is for $L = 2L_i$ and Fig. 4 for $L/2L_i = 3$.

There is an upper limit on A but not lower limit. The reason is as follows: As A increases, the minimum value of E (at $x = 0$) decreases and the maximum value of s increases. So if A is too large, the maximum s will be so large that we cannot neglect recombination, which becomes important when np approaches n_i^2, or s approaches 1. Fre-

quently recombination can be neglected over parts of the intrinsic region but not near the center, where the field is a minimum and the carrier concentration a maximum. Then (3.4) will represent the field distribution over that part of the region where recombination is unimportant. The correct solution will join onto the cubic as we move away from the center of the intrinsic region, which will no longer be at the $x = 0$ point on the cubic. In Section IV we solve the equations for the recombination region and show how the solution approaches the cubic. We will show that, as A increases, the distance from the center of the intrinsic region to the $x = 0$ point on the cubic also increases. The value $A = \frac{2}{3}$ corresponds to an infinitely long intrinsic region. For a larger A there exists no exact solution that could join onto the cubic as recombination becomes negligible. In Figs. 3 and 4 the $A = \frac{2}{3}$ curves join onto recombination solutions at values of E which are too low to show.

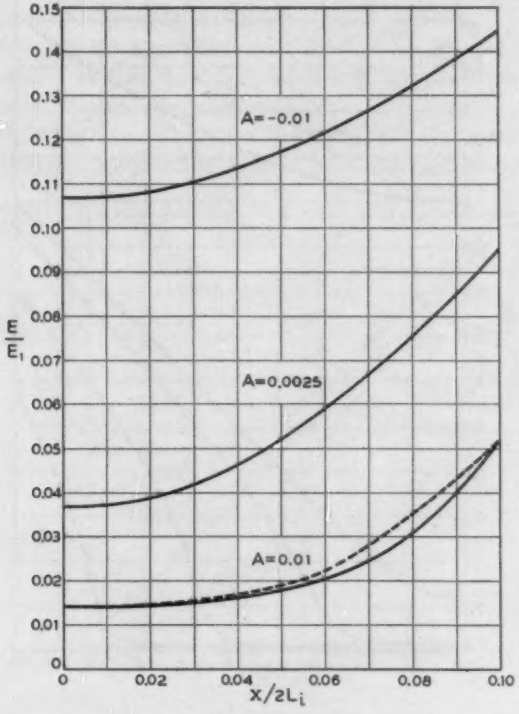


Fig. 2 — Field Distributions for $L = 0.2L_i$.

creases). The current generated in the intrinsic region is qg per unit volume. So the density of current from pairs generated in the intrinsic layer is $2Lqg = qn_iL/\tau$. Hence

$$I = I_s + \frac{qn_iL}{\tau}$$

In what follows we shall assume that I_s is negligibly small compared to I. Then

$$I = \left(\frac{qn_i}{\tau}\right)L = \left(\frac{qn_iD}{L_i}\right)\frac{L}{L_i} = \left(\frac{\sigma}{2L_i}\,\frac{kT}{q}\right)\frac{L}{L_i}$$

Thus I is L/L_i times a characteristic current equal to (1) the diffusion current produced by a gradient n_i/L_i or (2) the drift current produced by a field that gives the voltage drop kT/q in two diffusion lengths in normal intrinsic material. In germanium this characteristic current is about 5 milliamperes per cm^2.

That the current I is proportional to L and independent of voltage follows from the neglect of recombination. When recombination is small compared to generation, then the current has reached its maximum, or saturation, value. All the carriers generated in the intrinsic region are swept out before recombining. It will sometimes be convenient to take σE_1 as the unit of current. From the above and (2.8)

$$\frac{I}{\sigma E_1} = \frac{\sqrt{2}\mathfrak{L}L}{(2L_i)^2} \tag{3.5}$$

In germanium σE_1 is about 7 amperes per cm^2. In general we will be dealing with currents that are small compared to this. For example, if L_i is 1 mm, we would have to sweep out an intrinsic region 3 meters long in order to get a current this large. If we take E_1 as the unit field, σE_1 as the unit current and $2L_i$ as the unit length then the cubic becomes $E^2 - I/E = x^2 - A$.

For a given structure and temperature the field versus x curves form a one parameter family. A determines both the field distribution and the voltage. The voltage increases as A decreases. Fig. 2 is a plot of E/E_1 versus $x/2L_i$ for $L/2L_i = 0.1$ and several different values of A. Fig. 3 is for $L = 2L_i$ and Fig. 4 for $L/2L_i = 3$.

There is an upper limit on A but not lower limit. The reason is as follows: As A increases, the minimum value of E (at $x = 0$) decreases and the maximum value of s increases. So if A is too large, the maximum s will be so large that we cannot neglect recombination, which becomes important when np approaches n_i^2, or s approaches 1. Fre-

quently recombination can be neglected over parts of the intrinsic region but not near the center, where the field is a minimum and the carrier concentration a maximum. Then (3.4) will represent the field distribution over that part of the region where recombination is unimportant. The correct solution will join onto the cubic as we move away from the center of the intrinsic region, which will no longer be at the $x = 0$ point on the cubic. In Section IV we solve the equations for the recombination region and show how the solution approaches the cubic. We will show that, as A increases, the distance from the center of the intrinsic region to the $x = 0$ point on the cubic also increases. The value $A = \frac{2}{3}$ corresponds to an infinitely long intrinsic region. For a larger A there exists no exact solution that could join onto the cubic as recombination becomes negligible. In Figs. 3 and 4 the $A = \frac{2}{3}$ curves join onto recombination solutions at values of E which are too low to show.

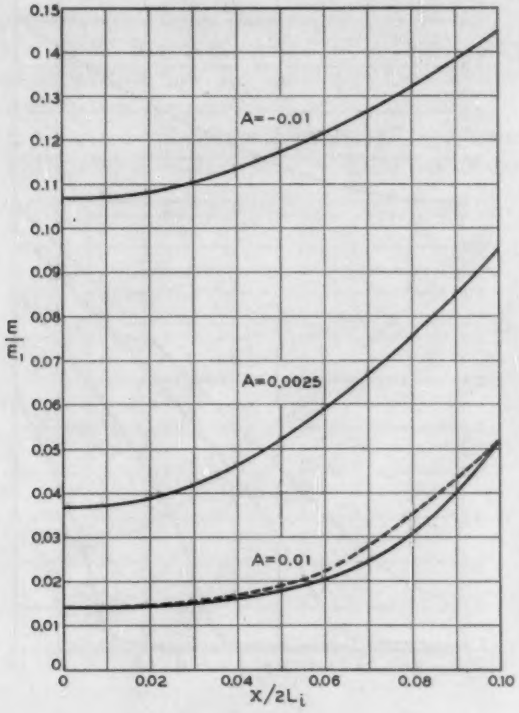


Fig. 2 — Field Distributions for $L = 0.2L_i$.

As A decreases and becomes negative the cubic approaches the form

$$E^2 = E_0^2 + E_1^2 \left(\frac{x}{2L_i}\right)^2 \tag{3.6}$$

where $E_0^2 = -AE_1^2$ is the minimum value of E^2. This form of the solution will be valid when the minimum E is large compared to (IE_1^2/σ). As E_0 increases, the voltage increases and the curve becomes flatter. This is because the increasing field sweeps the carriers out and reduces the space charge; so the drop in field decreases.

If (3.4) for E/E_1 versus $x/2L_i$ is extended to indefinitely large values of $x/2L_i$, it approaches the straight line of slope 1 going through the origin. Since E is always positive the curve is above this straight line at $x = 0$. If A is negative the curve is always above the straight line and always concave upward. If A is positive, the curve crosses the straight

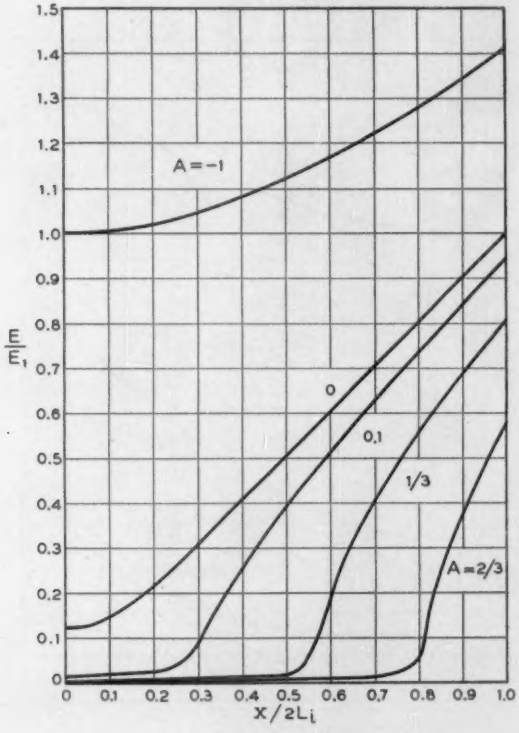


Fig. 3 — Field Distributions for $L = 2L_i$.

line at $E/E_1 = I/\sigma E_1 A$ and thereafter remains under it approaching it from below. For positive A the curvature, which is upward near the origin, changes to downward at about $x/2L_i = \sqrt{A}$.

The carrier concentrations n and p can be found from the E versus x curves with the aid of Poisson's equation $p - n = 1/a\, dE/dx$ and the definition $s = (n + p)/2n_i$ with $s = I/\sigma E$. These relations and (3.4) give

$$\frac{p-n}{p+n} = \frac{x}{L}\frac{1}{\left(1 + \frac{IE_1^2}{2\sigma E^3}\right)} \tag{3.7}$$

From (3.4) and (3.7) we may distinguish the following two regions on the cubic:

(1) When E^3/E_1^3 is smaller than $I/\sigma E_1$ (which as we have seen is usually smaller than unity), the E versus x curve is concave upward, the hole and electron concentrations are almost equal (charge neutrality) and the particle flow is by diffusion.

(2) When E^3/E_1^3 is greater than $I/\sigma E_1$, in general there is space charge and the particle flow, like the charge flow, is by drift. The curve is concave downward for positive A.

Figure 6, which we will discuss in Section IV, shows the field and carrier distributions for $L = 2L_i$ and $A = 0.665$ plotted on a logarithmic

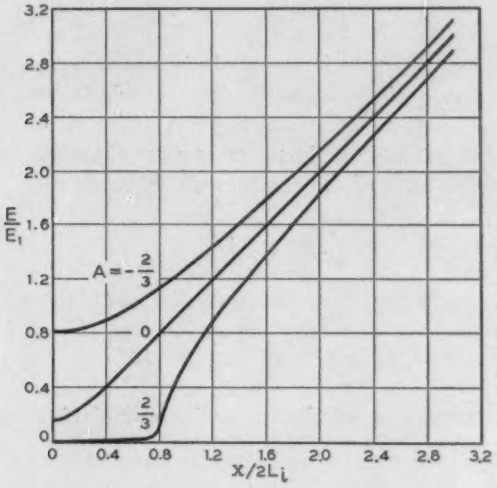


Fig. 4 — Field Distributions for $L = 6L_i$.

As A decreases and becomes negative the cubic approaches the form

$$E^2 = E_0^2 + E_1^2\left(\frac{x}{2L_i}\right)^2 \tag{3.6}$$

where $E_0^2 = -AE_1^2$ is the minimum value of E^2. This form of the solution will be valid when the minimum E is large compared to (IE_1^2/σ). As E_0 increases, the voltage increases and the curve becomes flatter. This is because the increasing field sweeps the carriers out and reduces the space charge; so the drop in field decreases.

If (3.4) for E/E_1 versus $x/2L_i$ is extended to indefinitely large values of $x/2L_i$, it approaches the straight line of slope 1 going through the origin. Since E is always positive the curve is above this straight line at $x = 0$. If A is negative the curve is always above the straight line and always concave upward. If A is positive, the curve crosses the straight

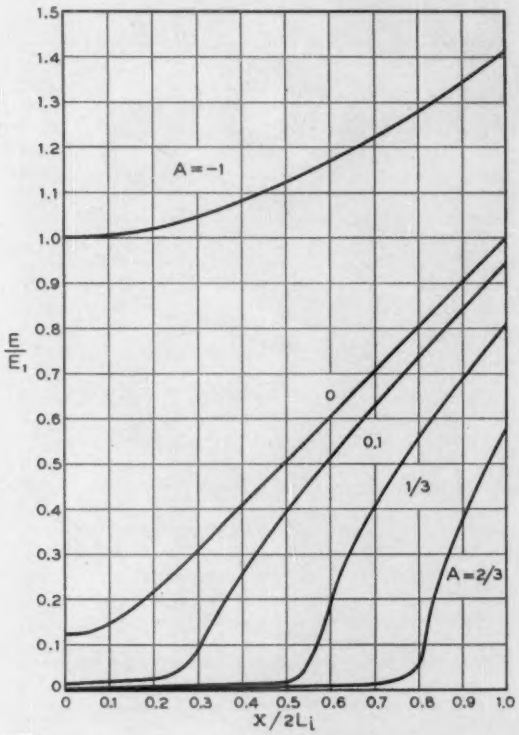


Fig. 3 — Field Distributions for $L = 2L_i$.

line at $E/E_1 = I/\sigma E_1 A$ and thereafter remains under it approaching it from below. For positive A the curvature, which is upward near the origin, changes to downward at about $x/2L_i = \sqrt{A}$.

The carrier concentrations n and p can be found from the E versus x curves with the aid of Poisson's equation $p - n = 1/a\, dE/dx$ and the definition $s = (n + p)/2n_i$ with $s = I/\sigma E$. These relations and (3.4) give

$$\frac{p - n}{p + n} = \frac{x}{L} \frac{1}{\left(1 + \frac{IE_1^2}{2\sigma E^3}\right)} \tag{3.7}$$

From (3.4) and (3.7) we may distinguish the following two regions on the cubic:

(1) When E^3/E_1^3 is smaller than $I/\sigma E_1$ (which as we have seen is usually smaller than unity), the E versus x curve is concave upward, the hole and electron concentrations are almost equal (charge neutrality) and the particle flow is by diffusion.

(2) When E^3/E_1^3 is greater than $I/\sigma E_1$, in general there is space charge and the particle flow, like the charge flow, is by drift. The curve is concave downward for positive A.

Figure 6, which we will discuss in Section IV, shows the field and carrier distributions for $L = 2L_i$ and $A = 0.665$ plotted on a logarithmic

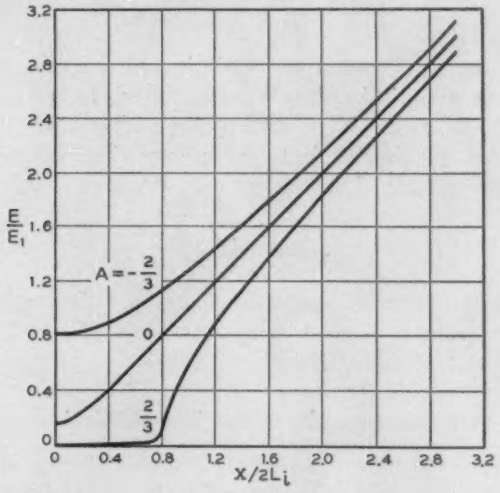


Fig. 4 — Field Distributions for $L = 6L_i$.

scale to show the behavior at low values of field and carrier density. In the region of no-recombination the field distribution is indistinguishable from that for $A = \frac{2}{3}$, which is plotted in Fig. 3 on a linear scale. In the region where recombination is important the solution is found from the assumption of charge neutrality as will be discussed in Section IV. The cubic and charge neutrality solutions are each shown dashed outside of their respective ranges of validity. For $A = 0.665$ the half length of the intrinsic region is $2.098 \times 2L_i$. Thus the length of the intrinsic region is more than twice the effective length $2L$ in which current is generated. The effective length will be discussed in more detail in Section IV and it will be shown that the effective length $2L$ of current generation is equal to the twice the distance from the IP junction to the minimum on the cubic. As explained earlier, it is convenient to take $x = 0$ at the minimum on the cubic.

Intrinsic-Extrinsic Junction Under Large Bias

Consider the limiting case of an intrinsic-extrinsic junction as the bias is increased and the space charge penetrates many diffusion lengths into the intrinsic material. Then the field distribution approaches the straight line $E/E_1 = x/2L_i$. This, by Poisson's equation, means that there is a constant charge density of N_i where

$$N_i = \frac{E_1}{2aL_i} = \frac{\sqrt{2}\mathfrak{L}}{L_i} n_i$$

Thus in the limit, the field in the intrinsic region approaches that in a completely swept extrinsic region having a fixed charge density of N_i. In germanium at room temperature N_i is about 4.10^{10} cm^{-3}. As the field approaches the limiting form, the voltage V approaches $E_1L^2/4L_i$. Thus the limiting form of the current voltage curve is

$$\frac{I}{\sigma E_1} = \frac{\mathfrak{L}}{L_i}\sqrt{\frac{V}{2E_1L}}$$

So in the limit the current varies as the square root of the voltage. Typical values for germanium at room temperature are $\sigma E_1 = 7$ amps cm^{-2}, $\mathfrak{L}/L_i = 10^{-3}$ and $2E_1L_i = 50$ volts.

Equivalent Generation Length for an Intrinsic-Extrinsic Junction

It should be noted that for an IP structure the current is the same as for an NIP structure with an infinite I region, or at least an I region that is long compared to the distance of penetration of the space charge.

Thus the equivalent length of current generation is $2L$ even though the current is actually being generated in an effective length L. The reason is that for an NIP structure the holes entering the right hand half of the I region were generated in the left hand side. For an IP structure the holes entering the space charge regions from the left were injected at the external left hand contact to the I region.

Applied Voltage

In all cases the voltage can be found from the area under the E versus x curve. In Figs. 2 to 4 the area under the curves gives the voltage accurately; recombination becomes important only where the field is so low as to have a negligible effect on the total voltage drop.

Correction of the Cubic

To conclude this section we consider the error introduced by using the assumption $I = \sigma sE$. For simplicity take E_1 as the unit field, $2L_i$ as the unit length and σE_1 as the unit current. Then the cubic becomes $E^3 - I/E = x^2 - A$. The corresponding exact solution is $E^2 - s = x^2 - A$ where the relation between s and E is given by equation (2.6) which in dimensionless form is

$$\mathfrak{L}^2 \frac{d^2E}{dx^2} = sE - I \tag{3.8}$$

where $\mathfrak{L}^2$ is of the order of 10^{-6}.

Let δE and δs represent the difference between the cubic and the correct solution at a given x. Assume that δE and its second derivative are small compared to E and its second derivative respectively. Then $\delta s = 2E\delta E$ and on the correct solution $sE - I = (2E^2 + I/E)\delta E$. So (3.8) becomes

$$\frac{\delta E}{E} = \left(\frac{\mathfrak{L}^2}{2E^3 + I}\right)\frac{d^2E}{dx^2} \tag{3.9}$$

To obtain a first approximation to $\delta E/E$ we use the cubic to evaluate d^2E/dx^2. It is convenient to express the results in terms of a dimensionless variable $z = E/I^{1/3}$, or if E and I are measured in conventional units $z = E(\sigma/E_1^2 I)^{1/3}$. Then (3.9) becomes

$$\frac{\delta E}{E} = \frac{1}{2}\left(\frac{L_i\mathfrak{L}}{L^2}\right)^{2/3}\left(\frac{z}{z^3 + \frac{1}{2}}\right)^2 + \left(\frac{x}{2L}\right)^2 \frac{z^3(1 - z^3)}{(z^3 + \frac{1}{2})^4} \tag{3.10}$$

when the lengths are in conventional units.

The first term has a maximum value of 0.35 $(L_i\mathfrak{L}/L^2)^{2/3}$ at $z = 0.6$ and the second term a maximum value of 0.18 at $z = 0.5$ and $x = L$.

The dashed curve in Fig. 2 for $A = .01$ is the corrected cubic. For the other curves in Fig. 2, the correction is smaller. For the curves in Figs. 3 and 4 the correction is too small to show.

Limits on the Solution

We now show that δE as derived above is not only a first approximation but also upper limit on the correction necessary to take account of charge diffusion. That is, an exact solution to (3.8) lies between the cubic and the corrected cubic.

Consider the region where the second derivative of E is positive so that the perturbed curve lies above the cubic as in Fig. 2. On the cubic we have $sE - I = 0$. As we move upward from the cubic and toward the dashed curve, $sE - I$ increases. The value of $sE - I$ on the dashed curve just equals the value of $\mathfrak{L}^2\, d^2E/dx^2$ on the cubic. However, the dashed curve has a smaller second derivative than the cubic. Thus in moving upward from the cubic toward the dashed curve $sE - I$ increases from zero and $\mathfrak{L}^2\, d^2E/dx^2$, which is positive, decreases; on the dashed curve $sE - I$ is actually greater. Therefore there is a curve lying just under the dashed curve where the two sides of (3.8) are equal. The same argument applied to the region where the second derivative is negative shows that the equation is satisfied by a curve lying just above the first perturbation of the cubic. Where the curvature changes sign, the cubic is correct.

It should be emphasized again that the neglect of the diffusion term in the current is justified only for the ideal case of equal hole and electron mobilities. For unequal mobilities both drift and diffusion will be important in current flow. However, as we will discuss in section 5, we can simplify the problem of unequal mobilities by defining a fictitious current that has the same form as I in (2.6) and (3.8).

IV. RECOMBINATION

As discussed in Section III, when the voltage for a given current is reduced, s increases and near $x = 0$ becomes comparable to unity. Then recombination becomes important and the cubic solution breaks down, or rather joins onto a solution that takes account of recombination. When recombination is important the center x_i of the intrinsic region is no longer at the $x = 0$ point on the cubic but to the left of it. That is, if we want the same current with continually decreasing voltage, we even-

tually get to the point where a longer intrinsic region is required. Finally for a given current we reach a minimum voltage which corresponds to an infinite length of intrinsic region. Another way of saying this is that, when recombination becomes important, the length L defined in terms of the current by $I = qg2L = qn_i/\tau L$ is no longer the half length of the intrinsic region.

Equivalent Generation Length

We shall continue to define L by $I = qn_i/\tau L$. Thus L is an equivalent, or effective, half length of current generation and not the half length of the intrinsic region. By definition L is the length such that the amount of generation alone in the length L is equal to the net amount of generation (generation minus recombination) in the total half length of the intrinsic region. Hence

$$gL = \int_{x_i}^{x_p} (g - r)\, dx \tag{4.1}$$

where x_i is at the center of the intrinsic region and x_p at the IP junction. We shall for the most part deal with reverse biases of at least a few kT/q, in which case recombination is negligible at the junctions. Then the exact solution becomes the no-recombination solution before reaching the junctions. We shall continue to take $x = 0$ at the point $dE/dx = ds/dx = 0$ on the no-recombination solution which the exact solution approaches as recombination becomes negligible.

Simplifying Assumptions

The general differential equation with recombination will be the same as for no-recombination except that $g - r$ replaces g. From (3.1) and (3.2)

$$\frac{d^2}{dx^2}\left(\frac{E^2}{E_1^2} - s\right) = \frac{1}{2L_i^2}\left(1 - \frac{r}{g}\right) \tag{4.2}$$

From (2.12) and (2.13) and Poisson's equation

$$\frac{r}{g} = \frac{np}{n_i^2} = \left(\frac{n + p}{2n_i}\right)^2 - \frac{(n - p)^2}{(2n_i)} = s^2 - 2\left(\frac{\mathfrak{L}}{E_1}\frac{dE}{dx}\right)^2 \tag{4.3}$$

The following analysis will be based on the assumption of charge neutrality. That is we neglect terms in $p - n$ in comparison with those in $p + n$. In particular this means:

(1) The charge flows by drift so $I = \sigma sE$. This is the same assumption

made in the no-recombination case. It will be an even better approximation in the recombination region, where the second derivative of E is less.

(2) The particle flow is by diffusion. That is, E^2/E_1^2 can be neglected in comparison with s.

(3) The ratio of recombination rate r to generation rate g is proportional to $g - r$; that is $g - r = g(1 - s^2)$.

All of these simplifying assumptions can be justified by substituting the resulting solution into the original expressions and showing that the neglected terms are small when recombination is important. If the solution is substituted into (4.3) and (2.6) the neglected terms will turn out to be negligible — and therefore assumptions (1) and (3), justified — when s^2 is large compared to $\mathfrak{L}/L_i$. Assumption (2) follows from (1) and the fact that $I/\sigma E_1$ is small compared to unity.

Assumptions (2) and (3) may also be justified by the discussion following (3.7) in the following way: Where recombination is important s must be near unity. So the cubic will begin to break down when $s = I/\sigma E$ becomes near to unity, or when E approaches I/σ. However, if E is approximately I/σ then $\sigma E^3/IE_1^2$ is approximately $(I/\sigma E_1)^2$, which, as we saw in the Section III, is small compared to unity in practical cases. Thus recombination becomes important and the solution joins onto the cubic in the range where E^3/E_1^3 is small compared to $I/\sigma E_1$. In this range the particle flow is by diffusion and $p - n$ is small compared to $p + n$. As we move toward the center of the intrinsic region s increases and E and dE/dx decrease. Therefore, since assumptions (2) and (3) are good where the solution joins onto the cubic, they are good throughout the region where recombination is important.

The Recombination Solution

The differential equation (4.2) now takes the form

$$\frac{d^2s}{dx^2} = -\frac{(1 - s^2)}{2L_i^2} \tag{4.4}$$

The solution for s in the recombination range is seen to be the same for all values of the current. When s has been found E is found from $E = I/\sigma s$.

For small disturbances in normal carrier concentration, s is only slightly different from unity and (4.4) takes the familiar form

$$\frac{d^2}{dx^2}(1 - s) = \frac{1 - s}{L_i^2}$$

which says that the disturbance in carrier concentration varies exponentially as x/L_i.

Equation (4.4) can be integrated once to give

$$\left(\frac{ds}{dx}\right)^2 = \frac{1}{L_i^2}\left(s_0 - s - \frac{s_0^3 - s^3}{3}\right) \tag{4.5}$$

where s_0 is the value of s at the center of the intrinsic region where s is a maximum.

As recombination becomes unimportant, s^2 becomes small compared to unity and (4.5) approaches the form

$$\left(\frac{ds}{dx}\right)^2 = \frac{1}{L_i^2}\left[s_0\left(1 - \frac{s_0^2}{3}\right) - s\right] \tag{4.6}$$

and the solution joins onto the no-recombination solution.

Joining onto the Cubic.

We have seen that the solution joins onto the no recombination solution, in the region where particle flow is by diffusion so that the no recombination solution has the form $s = A - (x/2L_i)^2$. This is readily transformed to the form (4.6) with

$$A = s_0\left(1 - \frac{s_0^2}{3}\right) \tag{4.7}$$

Thus the one arbitrary parameter s_0 in the recombination solution determines the parameter A in the cubic that the recombination solution approaches. Since the maximum value of s_0 under reverse bias is unity, the maximum value of A is $\frac{2}{3}$. Negative values of A correspond to solutions where recombination is always negligible.

The s versus x Curve

To find s versus x we integrate (4.5). There is one constant of integration, which is fixed by the choice of $x = 0$. We have taken $x = 0$ at the point where $dE/dx = ds/dx = 0$ on the cubic. To make the recombination solution join the cubic we choose the constant of integration so that the recombination solution extrapolates to $s = 0$ at $(x/2L_i)^2 = A$. Then

$$\frac{x}{2L_i} = \sqrt{A} - \frac{\sqrt{3}}{2}\int_0^s \frac{ds}{\sqrt{3(s_0 - s) - (s_0^3 - s^3)}} \tag{4.8}$$

which can be expressed in terms of elliptic integrals.

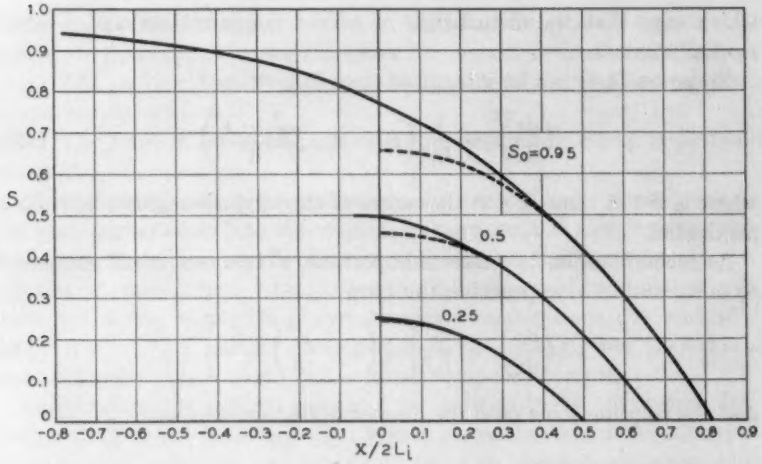


Fig. 5 — Variation of $s = p/n_i = n/n_i$ in the range where recombination is important.

Deep in an infinitely long intrinsic region the carrier densities approach their normal values $n = p = n_i$, or $s = 1$. Putting $s_0 = 1$ in (4.8), we find that as s approaches $s_0 = 1$, x becomes infinite. This will be the solution for a simple intrinsic-extrinsic junction. Fig. 5 is a plot of s versus x for various values of s_0. The dashed curves represent the corresponding no-recombination solution $s = A - (x/2L_i)^2$.

The IP Junction

It remains to find the position of the IP boundary. We now show that if recombination is unimportant at the junction, so that the solution joins onto a no-recombination solution, then the position of the junction is at $x = L$ where L is the effective length of current generation and $x = 0$ is the point where $dE/dx = ds/dx = 0$ on the no-recombination solution (which of course will not be valid at $x = 0$). The proof is as follows: From the definition (4.1) of L and (4.2)

$$L = \int_{x_i}^{x_p} (1 - r/g)\, dx = 2L_i^2 \int_{x_i}^{x_p} \frac{d^2}{dx^2}\left(\frac{E^2}{E_1^2} - s\right) dx = 2L_i^2 \left[\frac{d}{dx}\left(\frac{E^2}{E_1^2} - s\right)\right]_{x=x_p} \tag{4.9}$$

If the boundary comes where recombination is negligible so that $(E/E_1)^2 - s = (x/2L_i)^2 - A$, then (4.9) gives $x_p = L$. Physically

this means that the amount of recombination in the interval from $x = 0$ to $x = L$ is just equal to the excess amount of generation in the interval from the center of the intrinsic region to $x = 0$.

If the applied reverse bias is less than a few kT/q then recombination is important even at the junction and there is no joining onto a no-recombination solution. In this case (4.9) says that for a given choice of current (and hence of L) the boundary comes where

$$\frac{ds}{dx} = -\frac{L}{2L_i^2} \tag{4.10}$$

Example. Fig. 6, which we discussed briefly in Section III, is a plot of the field and carrier distributions for $L = 2L_i$ and $s_0 = 0.95$, for which $A = 0.665$. The hole and electron densities were found from (3.7) and $p + n = 2n_i s$ where s is found from Fig. 5. When s approaches s_0 (4.8) for x versus s takes the simple form

$$\frac{x - x_i}{2L_i} = \frac{s_0 - s}{1 - s_0^2} \tag{4.11}$$

This will be accurate when $s_0 - s$ is small compared to $1/s_0 - s_0$. We have used (4.11) to evaluate the s versus x curve beyond the range of the $s_0 = 0.95$ curve in Fig. 5.

It is seen that the recombination solution in Fig. 6 joins the cubic in the range where n and p are still almost equal.

Variable Lifetime.

Finally consider the general case where the variation in τ with carrier density cannot be neglected. Then, with $n = p = n_i s$, (2.15) becomes $\tau = \tau_0 + (\tau_i - \tau_0)s$ and L_i^2 in (4.4) is replaced by $D\tau[1 + (\tau_i/\tau_0 - 1)s]$ where τ_i/τ_0 is given by (2.17). The more general form of (4.4) can be solved graphically after one integration. The solution will join onto a cubic if $(\tau_i/\tau_0 - 1)s$ becomes small compared to unity before space charge becomes important. This will be so if $(\tau_i/\tau_0 - 1)^{3/2} I/\sigma E_1$ is small compared to unity.

V. THE JUNCTION SOLUTION

In this section we consider the solution near the junctions, where the assumption $I = \sigma s E$ breaks down. We shall deal with reverse biases of at least a few kT/q so that recombination is negligible at the junctions. The junction solution will therefore join onto the no-recombination solution. We shall use the cubic solution in the no recombination region.

Again it is convenient to use dimensionless variables with E_1 as the

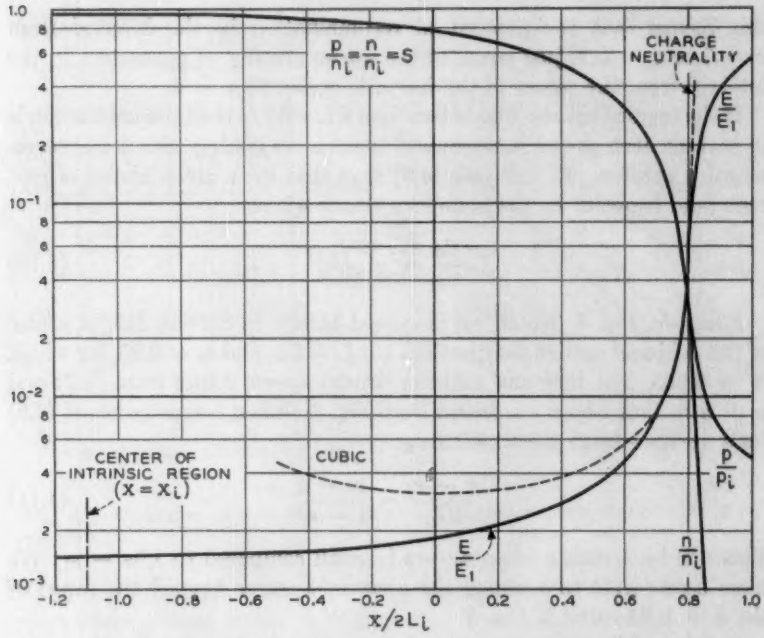


Fig. 6 — Field and carrier distributions for $L = 2L_i$ and $A = 0.665$ ($s_0 = 0.95$).

unit field, $2L_i$ as unit length and σE_1 as unit current. Then on the cubic $s = I/E$, and $E^2 - I/E = x^2 - A$. The current is related to L by $I = \sqrt{2}\mathfrak{L}L$ where the dimensionless $\mathfrak{L}$ is of the order of 10^{-3} for germanium at room temperature. Substituting the exact no-recombination solution $E^2 - s = x^2 - A$ into the solution (2.6), or (3.8), for the current gives the second order differential equation

$$\frac{d^2E}{dx^2} = \frac{1}{\mathfrak{L}^2}\left[E^3 - E(x^2 - A) - I\right] \tag{5.1}$$

for E as a function of x. The two boundary conditions are as follows: At $x = 0$, $dE/dx = 0$ by symmetry. At the IP junction the carrier concentration must rise and approach that in the normal P material. For a strongly extrinsic P region the normal hole concentration P is large compared to both n_i and the electron concentration. Thus s must increase and approach $P/2n_i \gg 1$ as we approach the P region. Clearly the cubic cannot satisfy this requirement. On the cubic the maximum value of s comes at $x = 0$ and is less than unity. As we approach the junc-

tion E increases so $s = I/E$ must decrease. Thus the correct solution must break away from the cubic near the junction.

Instability of the Solution

Equation (5.1) has two limiting forms and makes a rather abrupt transition between them. Over most of the intrinsic region, the quantity in brackets $[Es - I] = [E(E^2 - x^2 + A) - \mathrm{I}]$ almost vanishes. It differs from zero just enough that when multiplied by the very large factor $\mathfrak{L}^2 \approx 10^6$ it gives the correct second derivative of E. In Section III we derived an upper limit on the small deviation δE from the cubic required to satisfy the differential equation. If E deviates from the cubic by more than this small amount, then the second derivative of E becomes too large. This increases the deviation from the cubic, which further increases the second derivative and so on. E and s rapidly approach infinity in a short distance. This, of course, is the required behavior at the junction. The rapid increase in s makes it possible for s to approach $P/2n_i$.

In Section III we showed that there is a solution to the differential equation that lies within a small interval δE from the cubic. Suppose we try to solve (5.1) graphically or on a machine starting at $x = 0$. There are two boundary conditions: By symmetry $dE/dx = 0$ at $x = 0$. We choose for $E(0)$ a value somewhere in the interval $\delta E(0)$. The resulting solution will not long remain in the interval $\delta E(x)$. In fact there is only one choice of $E(0)$ for which the solution remains close to the cubic from $x = 0$ to $x = \infty$. For any other $E(0)$ the solution would remain close to the cubic for a certain distance and then abruptly become unstable and both E and s approach infinity. $E(0)$ must be so chosen that the solution becomes unstable and E and s become large at the junction. However it is impractical to set $E(0)$ on a machine with sufficient accuracy to insure that the solution will remain stable for a reasonable distance. A more practical procedure is to find a solution which holds near the junction and joins the cubic to a solution in the adjacent extrinsic region.

Zero Bias

It may be helpful to approach the junction solution by reviewing the simple case of an IP junction under zero bias. Both charge and particle flow vanish. The vanishing of particle flow means that in the intrinsic region $E^2 - s$ is constant, (2.7). Since $E = 0$ and $s = 1$ in the normal intrinsic material, it follows that $E^2 - s = 1$. With $I = 0$ the equation

for current becomes

$$\frac{d^2E}{dx^2} = \frac{sE}{\mathfrak{L}^2} = \frac{E^3 + E}{\mathfrak{L}^2} \tag{5.2}$$

This can be integrated at once. The boundary conditions are $dE/dx = 0$ when $E = 0$ and $E = E_j$ at $x = L$; the field E_j at the IP junction will be determined by joining the solutions for the I and P regions. The solution can best be expressed by parametric equations giving x and the potential V as functions of E.

$$L - x = \mathfrak{L} \int_E^{E_j} \frac{dE}{E\sqrt{1 + E^2/2}} = \mathfrak{L}\left[\operatorname{csch}^{-1}\frac{E}{\sqrt{2}} - \operatorname{csch}^{-1}\frac{E_j}{\sqrt{2}}\right] \tag{5.3}$$

$$V_j - V = \mathfrak{L} \int_0^E \frac{dE}{\sqrt{1 + E^2/2}} = \frac{2kT}{q}\left[\sinh^{-1}\frac{E_j}{\sqrt{2}} - \sinh^{-1}\frac{E}{\sqrt{2}}\right] \tag{5.4}$$

where we have used the relation between dimensionless quantities $\mathfrak{L} = \sqrt{2}kT/q$, which follows from (2.8) with $E_1 = 1$. It will be more convenient to express voltages in terms of kT/q rather than in terms of the unit voltage $2E_1L_i$; then the ratio qV/kT is independent of the units. For convenience we take the voltage as increasing in going toward the IP junction with $V = 0$ in the normal P material. The voltage V_j at the junction is found by joining solutions.

On the P side, let the excess acceptor density be P. Adding the term $-aP$ to the right hand side of Poisson's (2.1), and proceeding as before we have, instead of (2.5)

$$\frac{d}{dx}\left(\frac{E^2}{E_1^2} - s - s_p\frac{qV}{kT}\right) = J = 0$$

where $s_p = P/2n_i$. We shall assume that the P region is strongly extrinsic so that $n \ll p$. Then $s = s_p$ in the normal p material, where $E = V = 0$. Hence

$$E^2 - s = s_p\left(\frac{qV}{kT} - 1\right) \tag{5.5}$$

In the intrinsic material the corresponding solution is $E^2 - s = -1$. Since both E and s are continuous at the junction, $qV_j/kT = 1 - 1/s_p$ where $1/s_p$ can be neglected. Thus $E_j^2 = s_j = s_p \exp[-(qV_j/kT)] = s_p/e$ where $e = 2.72$ is the base of the natural logarithms.

Knowing E_j we can find the field and potential distributions in the intrinsic material from (5.3) and (5.4).

Reverse Bias

Now in the intrinsic region, $E^2 - s = x^2 - A$. Let E_c be the value of E at the junction as given by the cubic, and let $s_c = I/E_c$ be the corresponding value of s. Then at the junction $x^2 - A = E_c^2 - s_c$. In the P material equation (5.5) will still be a good approximation near the junction, where the additional terms arising from the flow will be negligible. Joining the solutions for the I and P regions and neglecting s_c in comparison with s_p gives

$$\frac{qV_j}{kT} = 1 + \frac{E_c^2}{s_p}$$

Again using $s_j = s_p \exp[-(qV_j/kT)]$ we have

$$E_j^2 = E_c^2 + s_p \exp[-(1 + E_c^2/s_p)] \tag{5.6}$$

In most practical cases E_c^2 will be small compared to $s_p = P/2n_i$ so E_j will be the same as for zero bias.

Junction Solution

We now consider an approximate solution that joins smoothly onto the cubic and has the required behavior at the junction. Let $x = x_0$ be the point where this solution is to join the cubic. Then in (5.1) x^2 must lie between x_0^2 and L^2. We can obtain two limiting forms of the solution by giving x the two constant values, x_0 and L respectively. It will be best to take $x = x_0$ since in practical cases the x^2 term is not important except near the point where the junction solution joins the cubic. In all cases the uncertainty due to taking $x^2 =$ constant can be estimated by comparing the solutions for $x = x_0$ and $x = L$.

With x^2 constant, (5.1) can easily be integrated. The two boundary conditions are (a) $E = E_j$ at $x = L$, where E_j is given by (5.6), and (b) to insure a smooth joining, the slope at $x = x_0$ must be the same as that of the cubic, namely

$$\left(\frac{dE}{dx}\right)_0 = \frac{2x_0}{2E_0 + I/E_0^2} \tag{5.7}$$

The first integration of (5.1) with $x = x_0$ gives

$$\left(\frac{dE}{dx}\right)^2 = \left(\frac{dE}{dx}\right)_0^2 + \frac{2}{\mathcal{L}^2}\left[\frac{E^4}{4} - \frac{E^2}{2}(E_0^2 - I/E_0) - IE\right]_{E_0}^{E} \tag{5.8}$$

where (dE/dx) is given by (5.7) and $E_0^2 - I/E_0 = x_0^2 - A$. The E versus

x curve can now be found from (5.8) and

$$x - x_0 = \int_{E_0}^{E} \left(\frac{dE}{dx}\right)^{-1} dE$$
$$L - x = \int_{E}^{E_j} \left(\frac{dE}{dx}\right)^{-1} dE \tag{5.9}$$

In general we will be intrested in cases where the junction solution holds over a length $L - x_0$ that is small compared to L, so we can take $x_0 = L$ in (5.7). It will also be valid to let E_0 in (5.7) and (5.8) be the value E_c on the cubic at $x = L$. Putting $E_c = E_0$ in equation (5.6) then gives E_j in terms of E_0 and $s_p = P/2n_i$, where P is the majority carrier concentration in the extrinsic region. In what follows we shall use these approximations. It will be convenient to express $x_0 = L$ in (5.7) in terms of I using $I = \sqrt{2}L\mathfrak{L}$. We continue to use dimensionless quantities with E_1, $2L_i$ and σE_1 as the units of field, length and current respectively, and $2L_iE_1$ as the unit of voltage. In general however we can express voltages in terms of kT/q.

When E_0^3 is either large or small compared to I, the junction solution takes a simple form and the field and potential distributions can be found analytically. We next consider two approximations that hold in those two cases respectively. Relatively good agreement between the two solutions at $E_0^3 = I$ indicates that each solution may be used up to this point.

Case of E_0^3 Large Compared to I

From (5.7) to (5.9)

$$x - x_0 = \sqrt{2}\mathfrak{L} \int_{E_0}^{E} \left[\left(\frac{I}{E_0}\right)^2 + (E^2 - E_0^3)^2\right]^{-1/2} dE \tag{5.10}$$

This can be solved in the following two overlapping ranges where the integrand has a simple form:

Range 1. Here $E - E_0$ is small compared to $2E_0$, so (5.10) becomes

$$x - x_0 = \frac{\mathfrak{L}}{\sqrt{2E_0}} \sinh^{-1}\left[(E - E_0)\frac{2E_0^3}{I}\right] \tag{5.11}$$

Since E and E_0 are almost equal, we have for the voltage drop in this range

$$V - V_0 = E_0(x - x_0) \tag{5.12}$$

Range 2. Here $E^2 - E_0^2$ is large compared to $2(\mathfrak{L}L/E_0)^2$, so (5.10) gives

$$L - x = \sqrt{2}\mathfrak{L} \int_E^{E_j} \frac{dE}{E^2 - E_0^2} = \frac{\sqrt{2}\mathfrak{L}}{E_0} \left(\text{ctnh}^{-1} \frac{E}{E_0} - \text{ctnh}^{-1} \frac{E_j}{E_0} \right) \quad (5.13)$$

From $E_0^3 \gg I$ it follows that Ranges 1 and 2 overlap. By joining the two solutions in the overlap region, the solution in Range 2 can be written as

$$x - x_0 = \frac{\mathfrak{L}}{\sqrt{2}\, E_0} \ell n \left[\frac{8E_0^3}{I} \frac{E - E_0}{E + E_0} \right] \quad (5.14)$$

Putting $E = E_j$ in (5.14) gives the distance over which the junction solution holds. In general we will be interested in cases where E_j is large compared to E_0 so (5.14) becomes

$$\frac{L - x_0}{l} = \frac{3}{2} \frac{\ell n\, 2z_0}{z_0} \quad (5.15)$$

where $l = \sqrt{2}\mathfrak{L}/I^{1/3}$ and as before $z_0 = E_0/I^{1/3}$. In conventional units

$$l = 2L \left(\frac{\mathfrak{L}L_i}{L^2} \right)^{2/3} \quad (5.16)$$

Fig. 7 is a plot of $(L - x_0)/l$ versus z_0. In germanium at room temperature $\mathfrak{L}L_i$ will be around 10^{-5} cm. Thus the junction solution will hold over a region that is small compared to L if L is large compared to 3×10^{-3} cm.

Again it is convenient to use the relation $\mathfrak{L} = \sqrt{2}kT/q$ to express the voltage in terms of kT/q.

$$V_j - V = \int_E^{E_j} E \left(\frac{dE}{dx} \right)^{-1} dE = \frac{kT}{q} \ell n \left(\frac{E_j^2 - E_0^2}{E^2 - E_0^2} \right) \quad (5.17)$$

By joining the two solutions in the overlap region, the voltage in Range 2 can be expressed as

$$V - V_0 = \frac{kT}{q} \ell n \left(\frac{2E_0}{I} \right) (E^2 - E_0^2) \quad (5.18)$$

Setting $V = V_j$ and $E = E_j$ in (5.18) gives the total voltage drop in the region where the junction solution holds. Let ΔV be the difference between $V_j - V_0$ and the built in voltage drop at the junction. Then substituting (5.6) with $E_c = E_0$ into (5.18) and subtracting the built in drop we have for ΔV,

$$\Delta V = \frac{kT}{q} \left[\ell n \frac{E_0}{I} - \frac{E_0^2}{s_p} \right] \quad (5.19)$$

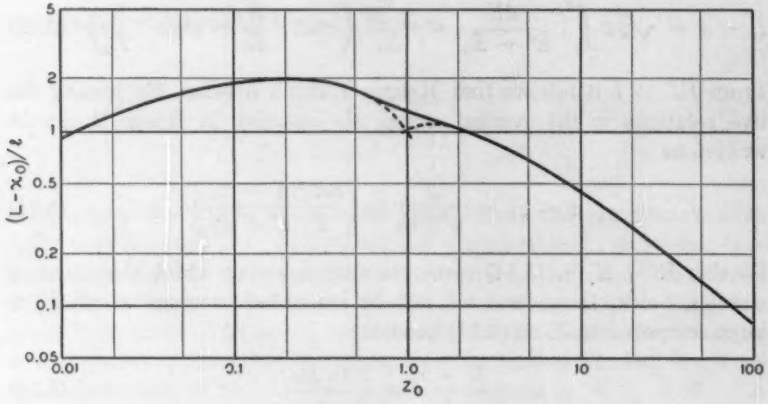


Fig. 7 — Variation of $(L - x_0)/\ell$ with z_0.

I/E_0 is equal to the value of s on the cubic at $x = L$. For positive values of A the maximum value of E_0/I is $L/I = 1/\sqrt{2}\mathfrak{L}$ as can be seen from the cubic. In germanium at room temperature $\mathfrak{L}$ is about 10^{-3} (for $2L_i$ = unit length) so the reverse bias produces an additional voltage drop in the junction region equal to about $7kT/q$. For negative values of A the additional voltage drop near the junction would be higher.

Comparing (5.3) and (5.13) we see that the junction solution reduces to the zero bias solution when E^2 is large compared to $E_0^2 + 2$. In this case both solutions have the simple form

$$L - x = \sqrt{2}\mathfrak{L}\left(\frac{1}{E} - \frac{1}{E_j}\right) \tag{5.20}$$

and

$$V_j - V = \frac{2kT}{q}\,\ell n\,\frac{E_j}{E} \tag{5.21}$$

Case of E_0^3 Small Compared to I

Now from (5.7) and (5.8) with $x_0 = L = I\sqrt{2}\mathfrak{L}$ we have

$$\begin{aligned}\left(\frac{dE}{dx}\right)_0^2 &= \left(\frac{2LE_0^2}{I}\right)^2 \\ \left(\frac{dE}{dx}\right)^2 &= \frac{1}{\mathfrak{L}^2}\left[2E_0^4 + (E - E_0)^2\left(\frac{E^2}{2} + \frac{I}{E_0}\right)\right]\end{aligned} \tag{5.22}$$

Again there are two overlapping ranges where the solution has a simple form:

Range 1. Here E^2 is small compared to $2I/E_0$. This will be so even when E becomes large compared to E_0. Setting $c_1^2 = 2E_0^5/I$ and $y = E - E_0$ in equation (5.22) and integrating gives

$$x - x_0 = \mathfrak{L}\sqrt{\frac{E_0}{I}}\int_0^{E-E_0}\frac{dy}{\sqrt{c_1^2 + y^2}} = \mathfrak{L}\sqrt{\frac{E_0}{I}}\sinh^{-1}\left(\frac{E - E_0}{c_1}\right) \tag{5.23}$$

and

$$V - V_0 = \frac{kT}{q}\sqrt{\frac{2E_0}{I}}\left(\sqrt{c_1^2 + (E - E_0)^2} - c_1\right) + 2E_0(x - x_0) \tag{5.24}$$

Range 2. Here E is large compared to E_0. It follows from $E_0^2 \ll I$ that E is also large compared to c_1. Setting $c_2^2 = 2I/E_0$ we have

$$L - x = \sqrt{2}\mathfrak{L}\int_E^{E_j}\frac{dE}{E\sqrt{E^2 + c_2^2}} = \mathfrak{L}\sqrt{\frac{E_0}{I}}\left(\operatorname{csch}^{-1}\frac{E}{c_2} - \operatorname{csch}^{-1}\frac{E_j}{c_2}\right) \tag{5.25}$$

Joining (5.21) and (5.23) where they overlap we have in range (2)

$$x - x_0 = \mathfrak{L}\sqrt{\frac{E_0}{I}}\,\ell n\left(\frac{2I}{E_0^3}\right)\left[\frac{E}{c_2 + \sqrt{c_2^2 + E^2}}\right] \tag{5.26}$$

Putting $x = L$ and $E = E_j$ in (5.26) gives the length $L - x_0$ in which the junction solution holds. If E_j is large compared to c_2, then

$$\frac{L - x_0}{l} = \sqrt{\frac{z_0}{2}}\,\ell n\frac{4}{z_0^3} \tag{5.27}$$

where as before $z_0 = E_0/I^{1/3}$ and l is given by (5.16). Fig. 7 is a plot of $(L - x_0)/l$ versus z_0. The two approximations (5.15) and (5.27) for $z_0^3 \ll 1$ and $z_0^3 \gg I$ respectively are shown dashed. Both become inaccurate as they are extended toward $z_0 = 1$. The point at $z_0 = 1$ was obtained graphically. Each approximation is in error by about 28 per cent here. The error will decrease as each approximation is extended away from $z_0 = 1$ toward its range of validity.

The voltage in Range 2 is given by

$$V_j - V = \frac{2kT}{q}\left[\sinh^{-1}\frac{E_j}{c_2} - \sinh^{-1}\frac{E}{c_2}\right] \tag{5.28}$$

or again joining (5.28) to the solutions in Range 1 we have in Range 2

$$V - V_0 = \frac{2kT}{q}\sinh^{-1}\frac{E}{c_2} + 2E_0(x - x_0) \tag{5.29}$$

The total voltage drop in the junction can be found by setting $V = V_j$ and $E = E_j$ in (5.29). The term $2E_0(L - x_0)$ will be negligible. When E^2 is large compared to $c_2^2 + 2$ the junction solution reduces to the zero current solution as can be seen by comparing (5.3) and (5.25). Then the solution has the simple form (5.20) and (5.21). E_j will always be large compared to c_2. (E_j^2 is approximately s_p/e and $c_2^2 = 2s_0$ where s_0 is the value of s where the junction solution joins the cubic.) Thus the difference ΔV between $V_j - V_0$ and the built in voltage is

$$\Delta V = \frac{kT}{q}\ell n\frac{E_0}{I} \tag{5.30}$$

Example. Fig. 8 shows the field distribution near the IP junction for the case $L = 2L_i$ and $A = \frac{2}{3}$, for which the intrinsic region is infinitely long. The field distribution near the junction, however, will be indistinguishable from that for $A = 0.665$, or $s_0 = 0.95$, for which the intrinsic region is about twice the effective length of current generation. We have taken $E_j = 30$, which corresponds to an excess acceptor density $P = 4.7 \times 10^3\, n_i$ in the P region. Over the range where the junction solution holds the cubic gives an almost constant field $E = E_0 = E_c$. The junction solution goes from the cubic to the zero bias solution in a distance of the order of the Debye length. The sum of the built in voltage and the voltage derived from the cubic differ from the correct voltage by the order of $\pounds E_1$ or about kT/q. The total voltage is about $0.3\, E_1L_i$, which would be about 11 volts in germanium at room temperature.

VI. GENERAL CASE, UNEQUAL MOBILITIES

This Section deals with the general case where the ratio of the hole and electron mobilities is arbitrary. The procedure is similar to that used in the preceding Sections. Many of the results for $b = 1$ are useful in the present, more general, case. We shall deal first with the no-recombination case and again find that E is given by a cubic. However, the field distribution is no longer symmetrical and the coefficient of the I/E term in the cubic is a linear function of x instead of a constant. The differential equation for s in the recombination region remains un-

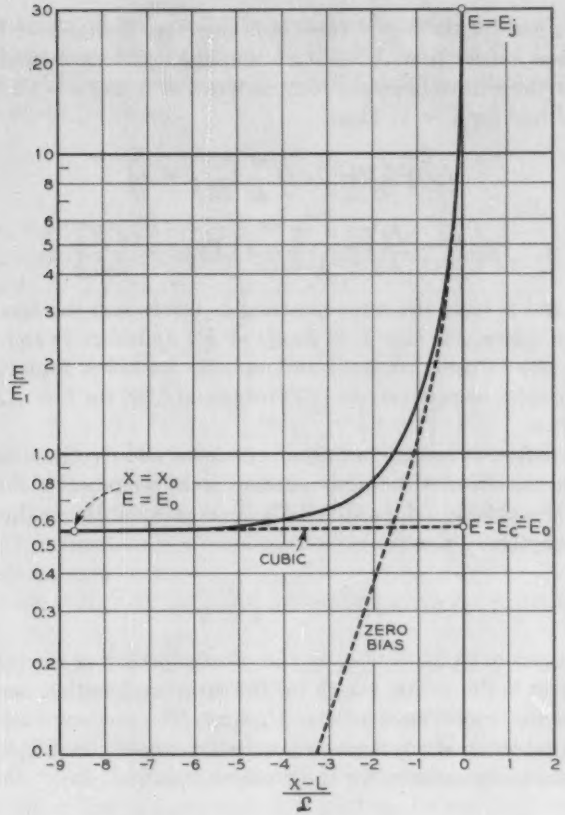


Fig. 8 — Field Distribution near the IP Junction for $L = 2L_i$ and $A = \frac{3}{2}$.

changed. It is no longer so that charge diffusion can be neglected except near the junctions. However, there is a linear combination of J_p and J_n in which the diffusion term is negligible except near the junctions.

Basic Relations

The equations are the two continuity (2.9) and Poisson's (2.1). The formulas for $g - r$ remain unchanged, since they involve only the statistics of recombination and are independent of mobility. The hole and electron currents are given by (2.2) with b arbitrary. Equation (2.2) for J_p in terms of E, p and n remains unchanged. Now J_n/b has the same

form as J_n had for the $b = 1$ case. It is therefore desirable to deal with the fictitious carrier flow $J_p + J_n/b$ and the fictitious current $q(J_p - J_n/b)$ since these have the same form in terms of E and $s = (n + p)/2n_i$ as J and I had for $b = 1$. Thus

$$J_p + \frac{J_n}{b} = 2n_i D \frac{d}{dx}\left(\frac{E^2}{E_1^2} - s\right) \tag{6.1}$$

$$q\left(J_p - \frac{J_n}{b}\right) = \frac{2}{1+b}\sigma\left[Es - \mathfrak{L}^2 \frac{d^2E}{dx^2}\right] \tag{6.2}$$

where E_1 and $\mathfrak{L}$ have the same meaning as before and the conductivity of intrinsic material is now $\sigma = qn_i\mu(1 + b)$. As before D and μ are respectively the diffusion constant and mobility for holes. Equations (6.1) and (6.2) reduce respectively to (2.7) for J and (2.6) for $I = q(J_p - J_n)$ where $b = 1$.

When the flow is by pure diffusion, the holes and electrons diffuse "in parallel" so the effective diffusion constant is the reciprocal of the average of the reciprocal hole and electron diffusion constants. Hence the effective diffusion length is given by

$$L_i^2 = D\tau \frac{2b}{1+b} \tag{6.3}$$

We continue to let $2L = I/qg$ be the effective length of current generation; again it is the actual length for the no recombination case. Let x_n and x_p be the coordinates of the NI and IP junctions respectively. Since the problem is not symmetrical we will not take $x = 0$ in the center of the intrinsic region even for the no-recombination case.

No-Recombination Case

Setting $r = 0$ we can immediately integrate the continuity equations

$$\frac{dJ_p}{dx} = \frac{dJ_n}{dx} = g$$

subject to the boundary conditions:

$$\begin{aligned} &\text{at the } NI \text{ junction, } x = x_n, \quad && J_p = 0, && J_n = -I/q \\ &\text{at the } IP \text{ junction, } x = x_p, && J_p = I/q, && J_n = 0 \end{aligned} \tag{6.4}$$

The result is $J_p = g(x - x_n)$ and $J_n = g(x - x_p)$. This agrees with $I = q(J_p - J_n) = 2qgL$ since $2L = x_p - x_n$ is the length of the intrinsic region, which, for no-recombination, is also the effective length of cur-

rent generation. It will be convenient to choose $x = 0$ so that $x_n = -x_p/b$. Then the origin is nearer to the NI junction for $b > 1$. Now from this and the boundary conditions (6.4) and $I = 2qgL$ we have the positions of the junctions:

$$\frac{x_p}{L} = \frac{2b}{1+b}, \qquad \frac{x_n}{L} = \frac{-2}{1+b} \tag{6.5}$$

As before, the junctions are at $x = \pm L$ for $b = 1$.

We can now find the fictitious carrier flow $J_p + J_n/b$ and the fictitious current $q(J_p - J_n/b)$ as functions of x.

$$J_p + \frac{J_n}{b} = \left(\frac{1+b}{b}\right) gx \tag{6.6}$$

$$q\left(J_p - \frac{J_n}{b}\right) = \frac{2I}{1+b}\left(1 + \frac{\beta x}{L}\right) \tag{6.7}$$

where the dimensionless parameter $\beta = (b^2 - 1)/4b$. Thus the fictitious current $q(J_p - J_n/b)$ is equal to the actual current times a linear function of x. This function is always positive and varies from a minimum of $1/b$ to a maximum of 1.

Combining (6.6) with (6.1) and integrating gives the equation

$$\frac{E^2}{E_1^2} - s = \left(\frac{x}{2L_i}\right)^2 - A \tag{6.8}$$

that we had before. Now, however, E is not a minimum at the same point where s is a maximum. As before, when recombination is negligible throughout all of the intrinsic region, A determines the voltage; and, when recombination is important over part of the region, A determines both the voltage and the length of the intrinsic region $x_p - x_n > 2L = I/gq$.

Combining (6.7) with (6.2) gives

$$I\left(1 + \frac{\beta x}{L}\right) = \sigma\left[Es - \mathfrak{L}^2 \frac{d^2E}{dx^2}\right] \tag{6.9}$$

which is similar to the previous (3.6) except that I is multiplied by the factor $1 + \beta x/L$, which varies from $1 + 1/b$ to $1 + b$. The same arguments used in Section V apply here and show that the second term in brackets (the diffusion term) can be neglected except near the junctions. In other words, although I is always part drift and part diffusion, $I(1 + \beta x/L)$ is approximately pure drift except at the junctions.

Eliminating s between (6.9) and (6.8) and neglecting the diffusion

term in (6.9) gives the cubic equation

$$\frac{E^2}{E_1^2} - \frac{I}{\sigma E}\left(1 + \frac{\beta x}{L}\right) = \left(\frac{x}{2L_i}\right)^2 - A \tag{6.10}$$

for the field distribution.

In germanium, where $b = 2.1$, $\beta = 0.406$, $x_p = 1.35L$ and $x_n = -0.65L$. The coefficient of $I/\sigma E_1$ therefore varies from 1.47 to 3.10, or by a factor of a little more than 2. This will introduce some asymmetry into the E versus x curve in the low field region where the fictitious carrier flow $J_p + J_n/b$ is by diffusion. It is evident that, as the voltage increases, the field versus x curve becomes increasingly symmetrical about the $x = 0$ point; so the effect of having $b \neq 1$ is simply to shift the field distribution along the x axis.

Recombination

The arguments of section 4 again apply. Where recombination is important, $n - p$ is small compared to $n + p$, so $g - r = g(1 - s^2)$. The diffusion term dominates in the fictitious particle flow $J_p + J_n/b$; that is, E^2/E_1^2 is small compared to s, so (6.1) becomes

$$J_p + \frac{J_n}{b} = -2n_i D \frac{ds}{dx}$$

The continuity equations give

$$\frac{d}{dx}\left(J_p + \frac{J_n}{b}\right) = \left(1 + \frac{1}{b}\right)(g - r) = \frac{n_i(1 + b)}{2\tau b}(1 - s^2)$$

So again we have

$$\frac{d^2 s}{dx^2} = -\frac{(1 - s^2)}{2L_i^2} \tag{6.11}$$

The solution joins the no recombination solution where $s = A - (x/2L_i)^2$. Therefore A is again related to s_0, the maximum s, by $A = s_0(1 - s_0^2/3)$ and the s versus x curve is given by (4.8) and is symmetrical about the point where s is a maximum. When the recombination solution joins onto no-recombination solutions, there will be a different no-recombination solution on each side of the recombination region. The junctions will be at the points x_p and x_n on the respective no-recombination solutions. The length of the intrinsic region will not be $x_p - x_n = 2L$ since the $x = 0$ points are different on the two no-recombination solutions and are separated by a region of maximum recombination.

To find E when s is known we express the current $I = q(J_p - J_n)$ in terms of s and E. Since $n - p$ is small compared to $n + p$, we set $n = p = sn_i$ in (2.2) and obtain

$$I = \sigma\left[sE - \frac{1-b}{1+b}\frac{kT}{q}\frac{ds}{dx}\right] \tag{6.12}$$

Thus the current contains both a drift and a diffusion term. This is to be expected for unequal mobilities. When holes and electrons have the same concentration gradient, the electrons, which have the higher diffusion constant, diffuse faster than the holes; hence the diffusion gives a net current. It is seen that in the recombination region the total carrier concentration has a symmetrical distribution about the point where it is a maximum but the field remains unsymmetrical.

Junction Solution

When $(E_0/E_1)^2$ is large compared to $I/\sigma E_1$ the junction solution is independent of b; so the solution obtained in Section V is valid. In all cases the junction solution can be found using the method of Section V. The effect of b will be small over most of the range where the junction solution holds because the concentration of one type of carrier will be negligible. To be exact, I in (5.8) should be multiplied by the factor $(1 + \beta x_0/L)$, which can be taken to be $(1 + b)/2b$ at the NI junction and $(1 + b)/2$ at the IP junction. Instead of equation (5.7) we have

$$\left[2E_0 + \frac{I}{E_0^2}\left(1 + \frac{\beta x_0}{L}\right)\right]\left(\frac{dE}{dx}\right)_0 = 2x + \frac{I}{E_0}\frac{\beta}{L} \tag{6.13}$$

as can be seen by differentiating (6.10) with $E_1 = 2L_i = \sigma = 1$.

VII. EFFECT OF FIXED CHARGE

This section will deal briefly with the case where there is some fixed charge but where the carrier charge cannot be neglected. For no recombination, the field distribution is given by a first order differential equation. Solutions in closed form are obtained for the case of pure drift flow. For recombination and charge neutrality the solution in Section IV is valid provided the fixed charge is small compared to n_i. We have seen that at large fields the E versus x curve becomes linear, corresponding to a fixed charge density of N_i where $N_i = \sqrt{2}n_i\mathfrak{L}/L_i$. Thus the fixed charge may have a dominant effect on the space charge while having a negligible effect on the solution in the range where recombination is important.

Let the density of fixed charge be $N = N_d - N_a$ = excess density of donors over acceptors. N may be either positive or negative. In what follows we shall assume that N is positive. So the structure is $N\nu P$ where ν means weakly doped n-type. Equations (2.2) for the hole and electron currents remain unchanged. Poisson's equation becomes

$$\frac{dE}{dx} = a(p - n + N) \tag{7.1}$$

We shall deal with the general case of arbitrary mobilities. As in Section VI it is convenient to deal with a fictitious current $q(J_p - J_n/b)$ and a fictitious particle flow $J_p + J_n/b$. The extra term in (7.1) drops out by differentiation when (7.1) is substituted into the equation for $J_p - J_n/b$ so (6.2) remains unchanged. However, instead of (6.1) we have

$$J_p + \frac{J_n}{b} = 2n_i D \frac{d}{dx}\left(\frac{E^2}{E_1^2} - s\right) - \mu NE \tag{7.2}$$

So the fictitious particle flow is no longer the gradient of a potential involving only E and s.

No Recombination

As in Section VI the continuity equations can be immediately integrated to give (6.6) and (6.7). Again I is given by (6.9) where the diffusion term on the right can be neglected except at the junctions; so again we have $\sigma sE = I(1 + \beta x/L)$. Substituting this into (7.2) and combining (7.2) and (6.6) gives a first order differential equation for E versus x. It is convenient to again use dimensionless quantities with E_1, $2L_i$ and σE_1 as the units of field, length and current respectively. Then the differential equation becomes

$$\frac{d}{dx}\left[E^2 - \frac{I}{E}\left(1 + \frac{\beta x}{L}\right)\right] = 2(x + \alpha E) \tag{7.3}$$

where

$$\alpha = \frac{N}{N_i}$$

and as before $N_i = \sqrt{2}n_i \mathfrak{L}/L_i$, which is around 4×10^{10} in germanium at room temperature. The solution of (7.3) contains one arbitrary constant (which corresponds to A in the $N = 0$ case). The lower limit on the constant is determined by the necessity of joining onto a recombination solution where s approached unity. The positions of x_n and x_p of the $N\nu$ and νP junctions respectively are given by (6.5).

In the region of low fields where E^3 is comparable to or less than I, (7.3) would have to be solved graphically or on a machine. At higher fields the equation is easily integrated as discussed below.

Case of Pure Drift

When the flow is entirely by drift, $E^3 \gg I$ and (7.3) becomes

$$\frac{dE}{dx} = \frac{x}{E} + \alpha \tag{7.4}$$

which is made integrable by the substitution $E = yx$. A family of solutions for positive E throughout the ν region is

$$(E - a_1x)^{a_1}(E + a_2x)^{a_2} = E_0^{a_1+a_2} \tag{7.5}$$

where $2a_1 = \sqrt{4 + \alpha^2} + \alpha$ and $2a_2 = \sqrt{4 + \alpha^2} - \alpha$ and E_0 is the value of E at $x = 0$. For an intrinsic region $N = \alpha = 0$ and (7.5) reduces to $E^2 = E_0^2 + x^2$, which is the same as (3.9) for negative A. Fig. 9 shows several curves for various values of E_0. These remain above, and at large distances approach, the asymptotic solutions $E = a_1x$ on the right of the origin and $E = -a_2x$ on the left. These curves differ from the corresponding curves for an intrinsic region in that the straight line asymptotes now have slopes of a_1 and $-a_2$ instead of ± 1. Toward the P side the slope is greater because the positive change qN of the excess donors is added to the charge of holes. Toward the N side of the ν region the slope is reduced because N compensates to some extent for the electron charge. As α increases and the ν region becomes more n type, the solution approaches that for a simple NP junction, where $E = \alpha x$ on the N side.

Another set of solutions of (7.4) are given by

$$(a_1x - E)^{a_1}(a_2x + E)^{a_2} = a_1^{a_1}a_2^{a_2}x_c^2 \tag{7.6}$$

Several of these are shown in Fig. 9. They remain below the linear asymptotes and go through zero field at $x = \pm x_c$. Actually these will join onto solutions of the more general equation (7.3) when E becomes small and the diffusion term becomes important.

Recombination. When the fixed charge density is small compared to the intrinsic hole and electron density the treatment of recombination in Section IV remains valid. The recombination solution joins onto a solution of (7.3) at small fields. When N is comparable to n_i the recombination solution is difficult even with the assumption of charge neutrality.

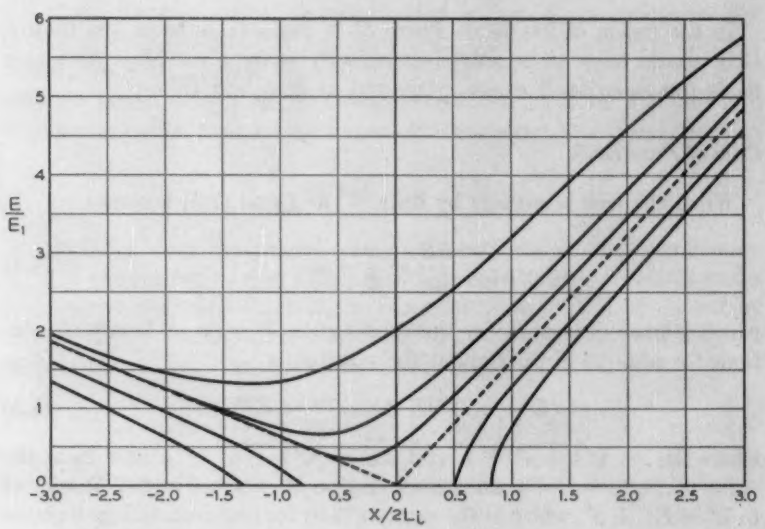


Fig. 9 — Field Distribution in the Range of Pure Drift for a fixed charge $N = N_i$, or $\alpha = 1$.

ACKNOWLEDGEMENTS

The author wishes to thank Miss M. M. Segrich for doing the extensive computations and plotting the curves, and Miss M. C. Gray for help with the calculations leading to Fig. 7.

APPENDIX A

Prim's Zero-Current Approximation

Prim's analysis is based on the assumption that the hole and electron currents are negligibly small differences between their drift and diffusion terms. Setting $J_p = J_n = 0$ then gives n and p as functions of the potential, which is found by substituting n and p into Poisson's equation and solving subject to the boundary conditions at the junctions. These conditions involve the applied bias and the majority carrier densities in the extrinsic regions. Since the current is assumed to vanish, the phenomena of carrier generation and recombination do not enter the problem and the results are independent of carrier mobility. The results will be exact when there is no applied voltage; the potential drop across the unit is then the built-in potential. In this appendix we use an internal consistency check to see for what values of applied bias the analysis

breaks down. First we find where the carrier concentration is in error by finding the bias at which the minimum drift current as calculated from $q\mu(n + p)E$ becomes equal to the total current, as found from the excess of generation over recombination in the intrinsic region. We then go on to find where the error in carrier concentration gives a sufficient error in space charge to affect the calculation of electric field. As we shall see, the zero-current approximation gives too low a carrier concentration in the interior of the intrinsic region. This will lead to serious errors in the field distribution only if the space charge of the carriers is important. When the bias is sufficiently high or the intrinsic region sufficiently narrow that the intrinsic region is swept so clean that the carrier space charge is, in fact, negligible, it will not matter that the calculated carrier density is too low, even by orders of magnitude. In such cases, the electric field is constant throughout most of the intrinsic region.

In the following we shall, for simplicity, take $b = 1$ and assume that the extrinsic regions are equally doped so that the problem is symmetrical.

Carrier Density

We now find where, on the zero current assumption, the drift current becomes equal to the total current. This involves knowing only the carrier concentrations and the field E_i in the center of the intrinsic region, where the drift current $q\mu(n + p)E_i$ is a minimum. By symmetry n and p are equal here and $n = p = n_i \exp(-qV_a/2kT)$ where V_a is the applied bias. The minimum field E_i is given by the total voltage drop V and the field penetration parameter η, which is the ratio of the minimum field to the average field. Thus $\eta = 2LE_i/V$ where $2L$ is the width of the intrinsic regions. The difference between V and V_a is the built-in voltage $(2kT/q)/\ell n(N/n_i)$ where N is the majority carrier concentration in the extrinsic regions. We now have for the drift current in the center of the intrinsic region

$$q\mu(n + p)E = q\eta D\left(\frac{qV}{kT}\right)\frac{n_i}{L}\exp\left(-\frac{qV_a}{2kT}\right) \tag{A1}$$

We next find the total current from the excess of generation over recombination in the intrinsic region. From the zero current assumption, $np = n_i^2 \exp(-qV_a/kT)$ is constant throughout the intrinsic region. Hence $g - r$ is constant. So the current $I = q(g - r)2L = qL(n_i^2 - np)/\tau n_i$ is

$$I = \frac{qLn_i}{\tau}\left[1 - \exp\left(-\frac{qV_a}{kT}\right)\right] \tag{A2}$$

Equating this to the drift current (A1) in the center of the intrinsic region gives

$$\left(\frac{L}{L_i}\right)^2 = \eta \frac{qV}{2kT} \operatorname{csch}\left(\frac{qV_a}{2kT}\right) \tag{A3}$$

The error in carrier concentration is less for narrower intrinsic regions and lower biases. Thus (A3) gives a curve of L versus V_a such that the zero current solution gives a good approximation to carrier concentration for points in the V_aL plane lying well below the curve. As expected, for zero bias, the solution is good for any value of L. However, for a bias of several kT/q, the solution for carrier concentration breaks down unless L is a very small fraction of a diffusion length.

Carrier Space Charge.

In Prim's analysis the carrier space charge is so low throughout most of the intrinsic region that the field remains approximately constant and equal to E_i. However there must be enough carriers present that the drift currents of holes and electrons can remove the carriers as fast as they are generated. In this section we ask where the space charge of the necessary carriers becomes large enough that its effect on the field can no longer be neglected. Let ΔE be the change in field due to the space charge in the intrinsic region (not counting the high field regions near the junctions). Unless ΔE is small compared to E_i the neglect of carrier space charge will not be justified. We shall find the ratio of ΔE to E_i.

If the field is to be approximately constant, then the hole and electron concentrations can easily be found from the hole and electrons currents. We shall deal with applied biases of at least a few kT/q, for which recombination is negligible and the total current $I = qg2L = qn_iL/\tau$. Since $g - r \doteq g$ is constant, the hole and electron currents are linear in x and, for constant field, are proportional to the hole and electron concentrations respectively. Thus the net space charge of the moving carriers $q(p - n)$ is proportional to x and varies from zero in the center of the intrinsic region to qp near the IP junction, where n is small compared to p and the current flows by hole drift, so $I = q\mu pE_i$. Thus the maximum charge is $I/\mu E_i$ and the total positive charge of the carriers on the P side of the center is $IL/2\mu E_i$. This gives a drop in field

$$\Delta E = \frac{aIL}{2q\mu E_i} = \frac{an_i}{2}\frac{kT}{qE_i}\frac{L^2}{L_i^2}$$

Dividing by $E_i = \eta V/2L$ gives

$$\frac{\Delta E}{E_i} = \frac{L^4}{\mathcal{L}^2 L_i^2}\left(\frac{kT}{\eta q V}\right)^2 \tag{A4}$$

Setting ΔE equal to some fraction, say 20 per cent of E_i gives a family of curves for V versus L with η as a parameter. Prim has plotted such curves in Fig. 11 of his paper. His curves will be good approximations when V for a given L and η lies above the V given by (A4).

Prim's results are expressed in terms of the parameters $U = qV/2kT$ and $\hat{L} = 2L/\mathcal{L}_e$ where $\mathcal{L}_e$ is the Debye length in the extrinsic material. $\mathcal{L}_e$ is given by the same formula as $\mathcal{L}$ except that N replaces n_i. Substituting these into (A4) and setting $\Delta E = E_i/5$ gives

$$\hat{L} = 3.57\,\frac{NL_i}{n_i\mathcal{L}}\,\eta U \tag{A5}$$

Prim's U versus $\hat{L}$ curves will be accurate up to the point where they intersect the corresponding curves from (A5). For germanium a reasonable value of $NL_i/n_i\mathcal{L}$ is about 10^6. This says that Prim's curves go bad at about $\hat{L} = 10^4$, which would be about 2.1×10^{-2} cm in germanium at 300°C.

Branching of the V versus L Curves

An effect which does not emerge from the zero-current analysis is that V may have several values for the same L and η. In other words the V versus L curve for given η will have more than one branch. Specifically, there will be a single V versus L curve up to a certain L at which the curve splits into three branches that diverge as L increases. This may be seen as follows: Consider an intrinsic region that is long compared to the diffusion length. Suppose a bias is applied that is low enough not to appreciably affect the space charge and potential drop at the junctions. A current will flow and a proportional, ohmic voltage drop will be developed across the intrinsic region. If the intrinsic region is long enough, this ohmic voltage may become large compared to the built-in voltage before the voltage drop at the junctions has changed appreciably. In this range the field penetration parameter will be rising from zero to about unity as V increases from the built-in voltage and approaches the ohmic voltage. As the voltage continues to increase, the space charge begins to penetrate the intrinsic region and a majority of the voltage drop comes in the space charge regions. Let L be the effective length of current generation. When L is larger than a diffusion

length but small compared to the length of the intrinsic region, then the voltage drop at the ends of the intrinsic region will be proportional to L^2 while the current, and consequently the minimum field, will be proportional to L. Thus η will be proportional to $1/L$ and will decrease as V increases and the region becomes more swept. Finally the two space charge regions meet; then η rises again with V and approaches unity. Hence, for a given η and length of intrinsic region, there will be three different values of V. For lower L the dip in the η versus V curve will be less, and there will be only one V for some values of η. Since η starts from zero at the built-in voltage and approaches unity for infinite voltage, there must be either one or three values of V for every η. Thus when the V versus L curve (or in Prim's notation the U versus $\hat{L}$ curve) branches, it branches at once into three curves. Prim's plot gives the upper branch in cases where all three are present.

A Medium Power Traveling-Wave Tube for 6,000-Mc Radio Relay

By J. P. LAICO, H. L. McDOWELL and C. R. MOSTER

(Manuscript received May 15, 1956)

This paper discusses a traveling-wave amplifier which gives 30 db of gain at 5 watts output in the 5,925- to 6,425-mc common carrier band. A description of the tube and detailed performance data are given.

TABLE OF CONTENTS

I. INTRODUCTION

During the past ten years traveling-wave tubes have received considerable attention in vacuum tube laboratories, both in this country and abroad. So far their use in operating systems has been somewhat limited, the most notable exceptions being in radio relay service in France, Great Britain, and Japan. However, it appears that sufficient progress in both tube and system design has been made so that traveling-wave tubes may see widespread application in the near future.

This paper describes an experimental helix type traveling-wave tube representative of a class which may see extensive use as a power amplifier in radio relay systems. The tube is designated as the Bell Laboratories type MI789. Stated briefly, the performance characteristics under nominal operating conditions are:

Frequency Range	5,925–6,425 mc
Power Output	5 watts
Gain at 5 watts output	31–35 db
Noise Figure	< 30 db

The tube is designed for use with waveguide input and output circuits. The input voltage standing wave ratio (VSWR) is less than 1.1 and the output VSWR is less than 1.4 over the 500-mc band when the tube is delivering 5 watts of output. Fig. 1 shows a photograph of an MI789 and of an experimental permanent-magnet focusing circuit.

In developing this tube we have endeavored to produce an amplifier which could be considered "practical" for use in a transcontinental radio relay system. Because such an application requires a high degree of reliability and refinement in performance, the tube was rather conservatively designed. This made it possible to obtain the desired gain and power output without difficulty. On the other hand, the contem-

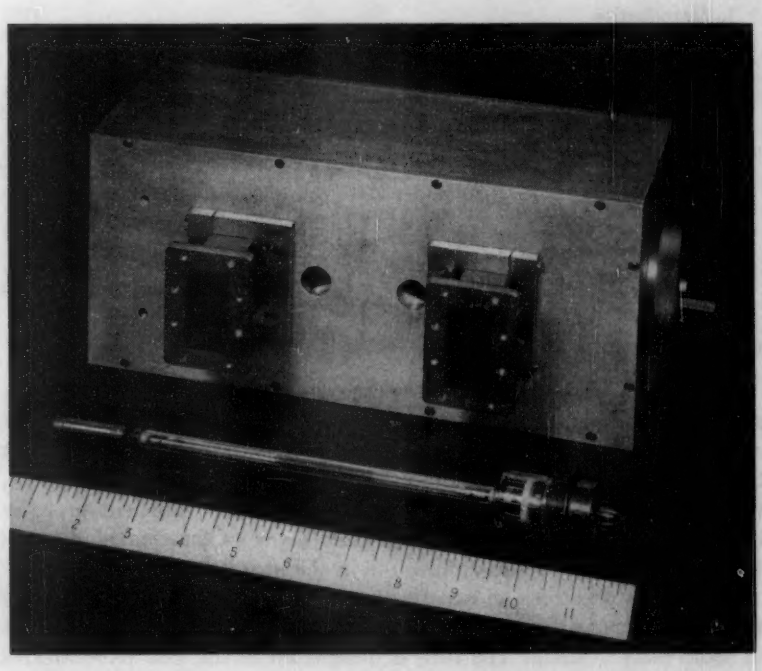


Fig. 1 — The M1789 traveling-wave tube and an experimental permanent magnetic circuit used to focus it. The circuit contains two specially shaped bar magnets between which the tube is mounted. The magnetic flux density obtained is 600 gauss, and the overall circuit weight is about 25 pounds.

plated system application made it necessary to investigate in detail the problems associated with band flatness, matching, noise output, certain signal distortions, reproducibility, and long life.

The solution of some of these problems required the development of a precisely constructed helix assembly in which the helix winding is bonded to ceramic support rods by glaze. Others required the initiation of a life test program. Early results indicate that life exceeding 10,000 hours can be obtained. This, in no small measure, is a result of a dc potential profile which minimizes the ion bombardment of the cathode. Since power consumed by focusing solenoids seriously degrades the overall efficiency of a traveling-wave amplifier, permanent magnet focusing circuits such as the one shown in Fig. 1 have been designed. Finally, to further improve efficiency, a collector which can be operated at about half the helix voltage was developed.

The major difficulties encountered in the course of the MI789 development were: excessive noise output, ripples in the gain-frequency characteristic, and lack of reproducibility of gain. There is evidence that a growing noise current wave on the electron stream was the source of the high noise output. This phenomenon has been observed by a number of experimenters but is not yet fully explained. By allowing a small amount of the magnetic focusing flux to link the cathode, the growing noise wave was eliminated, and the noise reduced to a reasonable level for a power amplifier. Reflections caused by slight non-uniformities in the helix pitch were the source of the gain ripples. Precise helix winding techniques reduced these reflections so that the ripples are now less than ± 0.1 db. The lack of reproducibility in gain was caused by variations in helix attenuation. Here, too, careful construction techniques alleviated the problem so that in a recent group of tubes the range of gain variation at five watts output was ± 2 db.

We have divided this paper into four main parts. The next section discusses some of the factors affecting the design of the traveling-wave tube. (We will henceforth use the abbreviation TWT.) Section III describes the tube itself. Certain performance data are included there when closely related to a particular portion of the tube. Section IV considers the rf performance in detail. There comparisons are made between the performance predicted from TWT theory and that actually observed. Finally Section V summarizes our life test experience.

This paper is written primarily for workers in the vacuum tube field and assumes knowledge of TWT theory. However, we believe that readers interested in TWT's from an application standpoint may also benefit from the discussion of the rf performance in Section IV. Much of that section can be understood without detailed knowledge of TWT's.

II. DESIGN CONSIDERATIONS

While TWT theory served as a general guide in the development of the MI789, a number of important tube parameters had to be determined either by experimentation or by judgement based on past experience. The most important of these were:

Saturation power output	12 watts
Mean helix diameter	90 mils
γa	$\sim$ 1.6
Magnetic flux density	600 gauss
Cathode current density	$\sim$ 200 ma/cm^2

These quantities and the requirement of 30-db gain at five watts output largely determined the TWT design.

The saturation output of 12 watts was found necessary to obtain the desired linearity at five watts output and the γa value of 1.6 to obtain the flattest frequency response over the desired band.

The choice of helix diameter and magnetic flux density represented a compromise. For the highest gain per unit length, best efficiency, and lowest operating voltage, a small helix diameter was called for. On the other hand, a large helix diameter was desirable in order to ease the problem of beam focusing and to facilitate the design of a light-weight permanent magnet focusing circuit. In particular, the design of such a circuit can be greatly simplified if the field strength required is less than the coercive force of available magnetic materials. This allows the use of straight bar magnets instead of much heavier horseshoe magnets. Moreover, the size and weight of the magnetic circuit is minimized by employing a high energy product material. These considerations led us to choose a flux density of 600 gauss, thereby permitting us to design a magnetic circuit using Alnico bar magnets.

To obtain long tube life we felt it desirable to limit the helix interception to about one per cent of the beam current. On the basis of past results we estimated that this could be done with a magnetic flux density 2.6 times the Brillouin value for a beam entirely filling the helix. With this restriction, Fig. 2 shows how the TWT design is affected by varying the helix diameter. A choice of 600 gauss is seen to result in a mean helix diameter of 90 mils.

In the selection of cathode current density, a compromise between long life and ease of focusing had to be made. To obtain long life, the current density should be minimized. However, this calls for a highly convergent gun which in turn complicates the focusing problem. We decided to use a sprayed oxide cathode operating at about 200 ma/cm^2. Experience with the Western Electric 416B microwave triode had shown

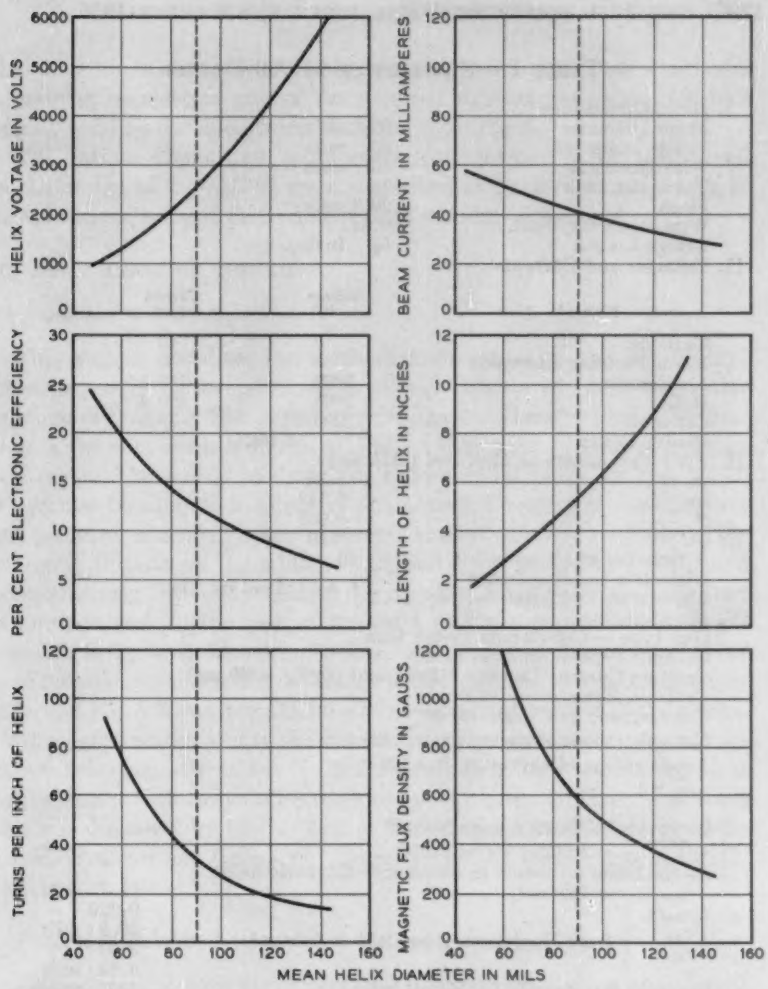


Fig. 2 — Alternate designs for the M1789. These curves are an estimate of how the TWT design would be affected by changing the helix diameter. They represent essentially a scaling of the M1789 design. In all cases the expected maximum power output is 12 watts and the low-level gain is 33 db. The line at 90 mils mean diameter in the curves represents the present M1789 design. In these calculations it was assumed that:

a. $\gamma a = 1.6$
b. power output $= 2.1\ CI_0V_0 = 12$ watts
c. the magnetic flux density is 2.6 times the Brillouin flux density for a beam entirely filling the helix.
d. the ratio of wire diameter to pitch is 0.34.
e. the dielectric loading factor is 0.79.
f. the ratio of effective beam diameter to mean helix diameter is 0.5.

TABLE I — SUMMARY OF M1789 DESIGN

I. Helix Dimensions

Mean Diameter	90	mils
Inside Diameter	80	mils
Wire Diameter	10	mils
Turns per Inch	34	
Pitch	29.4	mils
Wire Diameter/Pitch	0.34	
Active Length	5¼	inches

II. Voltages and Currents

Electrode	Voltage (volts)	Current (ma)
Cathode	0	40
Beam Forming Electrode	0	0
Accelerator	2600	< 0.1
Helix	2400	<0.4
Collector	1200	>39.5
Heater Power	6 watts	

III. TWT Parameters at Midband (6175 mc)

γa	1.58	
ka	0.148	
C	0.058	
QC	0.29	
N (number of λ's on helix)	30	
Dielectric Loading factor	0.79	As defined by Tien[3]
Impedance Reduction factor	0.4	As defined by Tien[3]

IV. Electron Gun

Gun type — Converging Pierce Gun
Cathode type — Sprayed oxide
Cathode Current Density 213 ma/cm² (for I_K = 40 ma)
Cathode diameter — 192 mils
Convergence half angle 12° 40′
Cathode radius of curvature ($\bar{r}_c$) 438 mils
Anode radius of curvature ($\bar{r}_a$) 190 mils
$\bar{r}_c/\bar{r}_a$ 2.3

Perveance 0.3×10^{-6} amps/volts$^{3/2}$
$\sqrt{V_A/T_K}$ = 1.61 for T_K = 720°C
At the beam minimum in absence of magnetic field:

r_{min} (from Pierce[10])	11.5	mils
r_{95}/r_c (from Danielson, Rosenfeld & Saloom[3])	0.220	
r_{95} (from Danielson, Rosenfeld & Saloom[3])	20.5	mils
r_c/σ (from Danielson, Rosenfeld & Saloom[3])	3.50	
σ (from Danielson, Rosenfeld & Saloom[3])	4.80	mils
Brillouin flux density for 80 mil helix ID	240	gauss
Actual focusing flux density required	600	gauss
Beam transmission from cathode to collector at 5 watts output	99%	

V. RF Performance

Frequency range	5925–6425 mc	
Saturation power output	12	watts
Nominal power output	5	watts
Gain at 5 watts	31–35	db
Noise figure	<30	db
Input VSWR	<1.1	impedance match to WR 159 waveguide
Output VSWR (at 5 watts)	<1.4	impedance match to WR 159 waveguide

For an explanation of symbols see page 1345.

that tube life in excess of 10,000 hours was possible with such a cathode. Moreover, an electron gun of the required convergence (about 13° half angle) could be designed using standard techniques.

The various dimensions, parameters, voltages and currents involved in the design of the MI789 are summarized in Table I. For the sake of completeness, some rf performance data are also included.

III. DESCRIPTION OF THE TUBE

3.1 *General Description*

This section describes the mechanical structure of the MI789 and presents some performance data closely associated with particular portions of the tube. The overall rf performance is reserved for consideration in the next section. In the MI789 we have tried to achieve a design which could be easily modified for experimental purposes and which would also be adapted to quantity production. To assist in obtaining low gas pressure, a rather "open" structure is used, thereby minimizing the pumping impedance. In addition, all parts are designed to withstand comparatively high temperatures during outgassing, both when the tube is pumped and, in the case of the helix and gun assemblies, during a vacuum firing treatment prior to final assembly. Fig. 3 shows an MI789 and its subassemblies. Fig. 4 shows a simplified drawing of the whole tube and Fig. 5 shows how the tube is mounted with respect to the permanent magnet circuit and to the waveguides. The permanent magnets are shown schematically in Fig. 5. In actual practice they are shaped so as to produce a uniform field between the pole pieces. The means of doing this was discussed by M. S. Glass at the Second Annual Meeting of the I.R.E. Professional Group on Electron Devices, Washington, D. C., October 26, 1956.

Control of Positive Ions

Our experience with previous TWT's has indicated that an improvement in life by as much as a factor of ten is obtained by arranging the dc potential profile so that positive ion bombardment of the cathode is minimized. This improvement has been observed even in tubes in which all reasonable steps have been taken toward minimizing the residual gas pressure. From Table I it is seen that the relative values of accelerator, helix, and collector voltage are arranged to drain positive ions formed in the helix region toward the collector. These ions are thereby kept from reaching the cathode. Spurious ion modulation which can result from accumulation of ions in the helix is also prevented.[1]

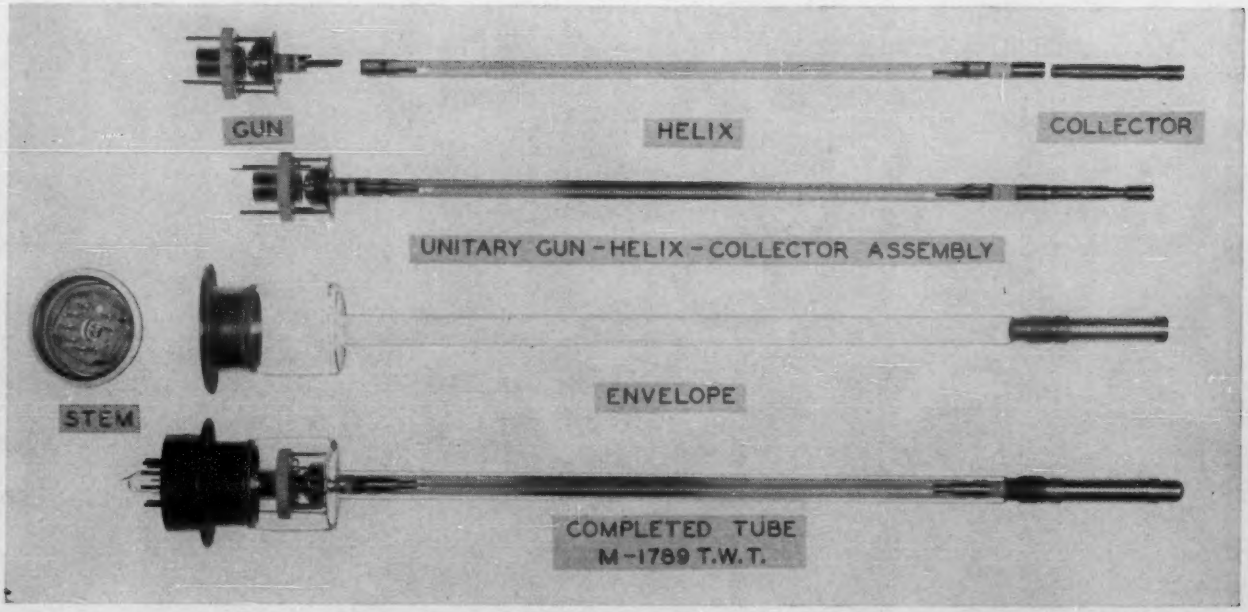
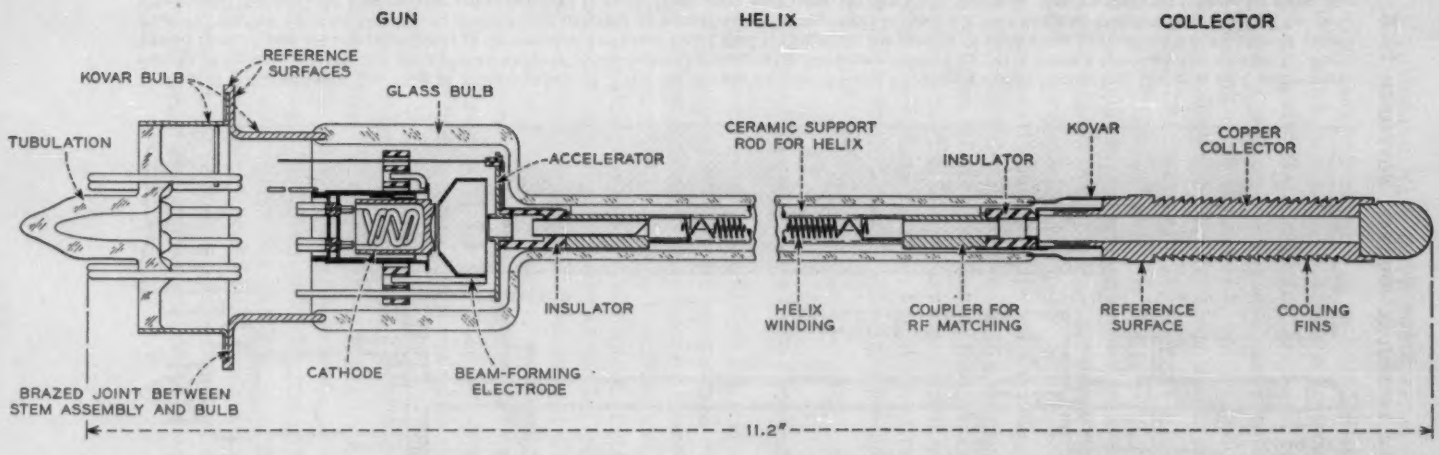


Fig. 3 — The M1789 TWT and its subassemblies. In assembling the tube, the gun is first connected to the helix-collector subassembly. This unit is then inserted into the envelope and an rf braze is made between the inner collector cylinder which is part of the helix assembly and the cooling fins which are part of the envelope. This braze extends for the entire length of the fins. Finally, the stem leads are connected to the gun and a second rf braze is made between the stem and the envelope.



1293

Fig. 4 — Simplified layout of the M1789. Detailed drawings of the various parts of the tube are shown in Figs. 8, 14 and 21.

The alignment surfaces are provided for mounting the TWT with respect to its associated magnetic circuit as is shown in Fig. 5. These surfaces are accurately concentric with the gun helix-axis. This is accomplished by shrinking the envelope in the helix region onto a precision mandrel and then grinding the surfaces concentric with this mandrel. The helix assembly is made a close fit inside of the glass envelope (less than two mils clearance), thereby making the helix axis concentric with the alignment surfaces. The gun is aligned with respect to the helix by telescoping cylinders which are held to a clearance of less than one mil.

The ceramic insulators at the ends of the helix provide rf isolation of the helix from the gun and collector. These insulators have a larger inside diameter than do adjacent metal parts to prevent them from charging as a result of electron bombardment. We have not observed any effects in the M1789 which could be attributed to charging of these ceramics.

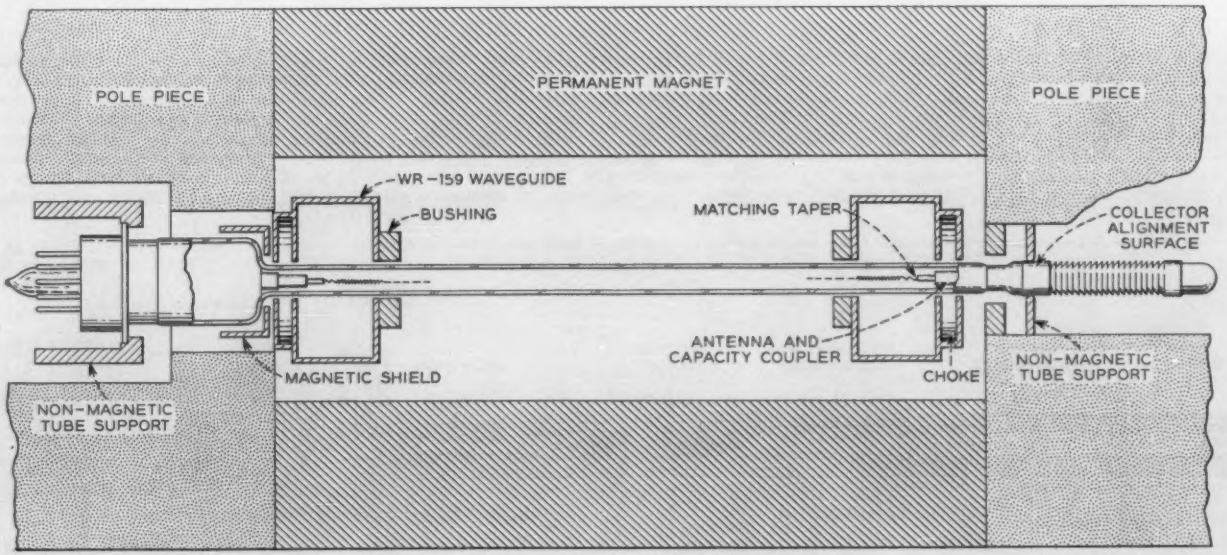


FIG. 5 — Schematic drawing of relationship of TWT to magnetic circuit and to waveguides. Mounting the gun and collector inside of holes through the pole pieces shields them from the magnetic focusing field. The helix passes through the center of the broad face of the waveguide. Energy is coupled between helix and waveguide by means of antennas and of matching tapers (see Fig. 14) at the ends of the helix. A shorting plunger is located in the waveguide about $\frac{1}{4}$ wavelength behind the TWT. The diameter of the cooling fins on the collector end is such that they can pass through the holes in the waveguide when the tube is inserted into the circuit. Forced air cooling of the collector is employed.

The effect that ions can have on cathode life was clearly demonstrated in a TWT which was in many aspects a prototype of the MI789. This tube operated with the accelerator, helix and collector at successively higher voltages, with consequent ion draining toward the cathode. Severe ion bombardment of the cathode brought about failure of most of these tubes in from 500 to 2,000 hours. In contrast to this the average life of the M1789 is in excess of 10,000 hours in spite of a cathode current density about twice that in the prototype tube. Moreover, failure of the M1789 comes about from exhaustion of coating material rather than as a result of ion bombardment. During the course of the work of the prototype tube, an experiment was performed to determine how much the ion bombardment would be affected by changing the potential difference between tube electrodes. In this experiment a small hole was drilled in the center of the cathode and an ion current monitoring electrode placed behind it. The ion monitor current was then investigated as a function of electrode voltages. Fig. 6 shows the results. We see that comparatively small potential differences are adequate to control the flow of positive ions.

3.2 *The Electron Gun and Electron Beam Focusing*

The electron gun used in the MI789 is a converging Pierce gun. The values of the gun parameters are summarized in Table I. Included are both the original parameters introduced by Pierce as well as those defined in a recent paper by Danielson, Rosenfeld and Saloom[2] in which the effects of thermal velocities are considered. Fig. 7 shows a drawing of the electrically significant contours of the MI789 gun. Fig. 8 shows the completed electron gun assembly. The method of constructing the gun is a modification of a procedure used in oscilloscope and television picture tubes. The electrodes are drawn parts made of molybdenum or, in the case of the cathode, of nickel. They are supported by rods which are in turn supported from a ceramic platform to which these rods are glazed. The whole gun structure is supported from the end of the helix by the helix connector detail. Since this part must operate at helix potential, it is insulated from the remainder of the gun by a ceramic cylinder which is glazed both to it and to the accelerator.

To obtain good focusing, the cathode must be accurately aligned with respect to the other electrodes. However, it must be omitted from the gun during the glazing process and during a subsequent vacuum outgassing because the cathode coating cannot withstand the temperatures involved. To insure proper placement of the cathode in the gun assembly

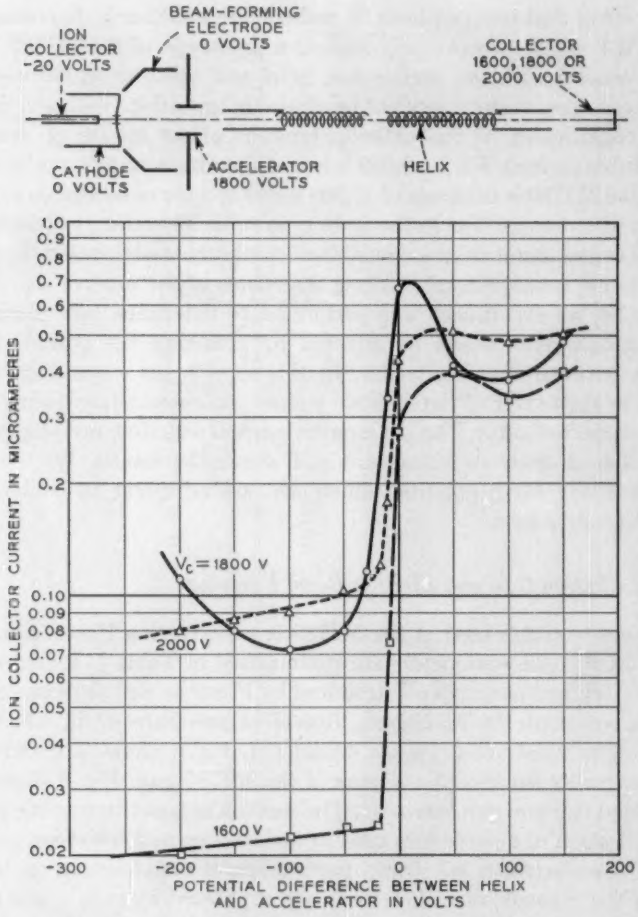


Fig. 6 — Effect of electrode voltages on ion bombardment of the cathode in a prototype of the M1789. In this experiment the helix voltage was varied while the positive ion current to a monitor electrode behind a hole in the cathode was measured. Curves are shown for the collector voltage greater than, equal to, and less than the accelerator voltage. During this experiment the accelerator voltage was held constant at 1800 volts with a resulting beam current of 40 ma. The experiment was performed on a continuously pumped system with the pressure maintained at 2×10^{-7} mm Hg. The helix ID was 80 mils, the cathode diameter 300 mils, and the cathode hole diameter 20 mils. These curves show that the ion bombardment of the cathode can be reduced by as much as a factor of 20 by properly arranging the voltage profile.

at a later stage, an alignment cylinder is included in the gun at the time of glazing (outer cathode alignment cylinder in Fig. 8). When the gun is ready to receive the cathode, the subassembly shown in Fig. 9 is slid into the outer alignment cylinder. The cathode to beam forming electrode spacing is set using a toolmakers microscope, and welds are made between the inner and outer alignment cylinders.

Initially, we thought that the cathode should be completely shielded from the magnetic field, and that the field should be introduced in the region between the accelerator and the point at which the beam would reach its minimum diameter in the absence of magnetic field. This ar-

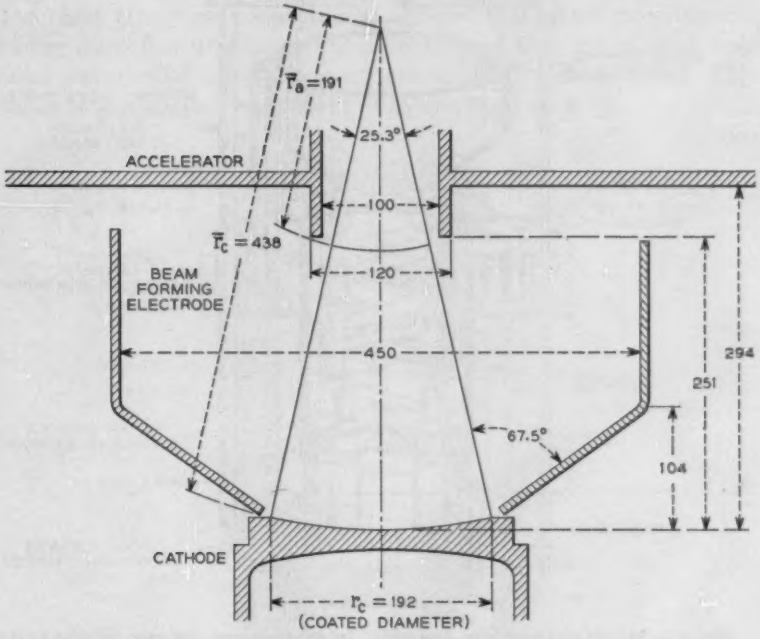


Fig. 7 — The electrically significant contours of the M1789 gun. All dimensions are in mils. These contours were determined using an electrolytic tank and following the procedure originated by Pierce. The measured potential at the beam boundary in the tank was made to match the calculated value within $\pm\frac{1}{4}$ per cent of the accelerator voltage to within 10 mils of the anode plane. The aperture in the accelerator was made sufficiently large so that substantially no beam current is intercepted on it. The significant parameters of this gun are:

P	$= 0.3 \times 10^{-6}$ amps/volts$^{3/2}$	r_a/σ	$= 3.50$	At the beam minimum in absence of magnetic field
$\bar{r}_c/r_a$	$= 2.30$	σ	$= 4.80$ mils	
θ	$= 12.67°$	r_{95}	$= 20.5$ mils	
$\sqrt{V_A/T_k}$	$= 1.61 (T_k = 720°C)$	J	$= 213$ ma/cm^2	

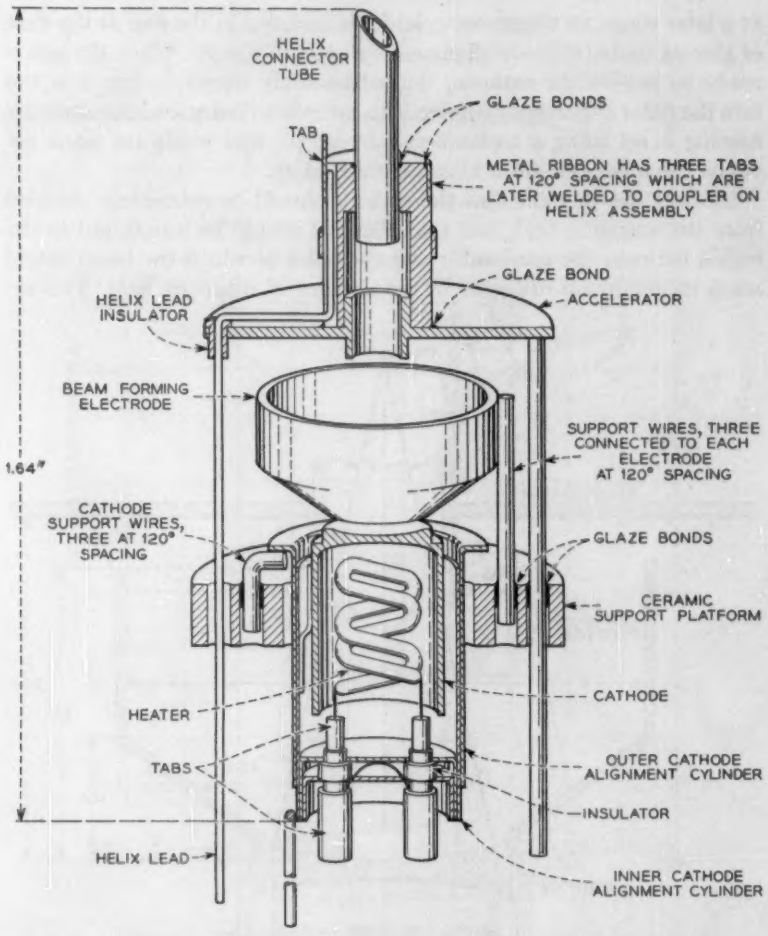


Fig. 8 — M1789 electron gun assembly. In constructing the gun, all the parts with the exception of the cathode, heater, and inner support cylinder are mounted on a mandrel which fixes their relative positions. Glass powder is applied to the areas where glazed joints are desired. The unit is then heated in forming gas (85% N_2, 15% H_2) to 1100°C to melt the glass and form the glazed bonds. With this technique the precision required for alignment and spacing of the electrodes resides entirely in the tools. The helix connector tube later slides into the coupler detail of Fig. 14 to align gun and helix assemblies. The inner and outer cathode alignment cylinders are welded together at two points at the end remote from the cathode. Optical comparator inspection shows that the significant dimensions of these guns are held to a tolerance of less than ±2 mils.

rangement did result in the best beam transmission to the collector. We later discovered, however, that the noise on the electron stream became extremely high when there was no magnetic flux at the cathode. This effect will be discussed further in Section IV. We found that by having a flux density of about 20 gauss at the cathode, the noise figure could be considerably reduced with the only penalty being a slight increase in interception on the helix. The penalty results from the fact that the flux linking the cathode causes a reduction in the angular velocity of the electrons in the helix region (from Busch's theorem), and this in turn diminishes the magnetic focusing force.

Fig. 10 shows the distribution of axial magnetic field in the gun region. The curve represents a compromise between that which gives best focusing (zero flux density at the cathode) and that which gives best noise performance (about 25 gauss flux density at the cathode). This flux density variation was arrived at by empirical methods.

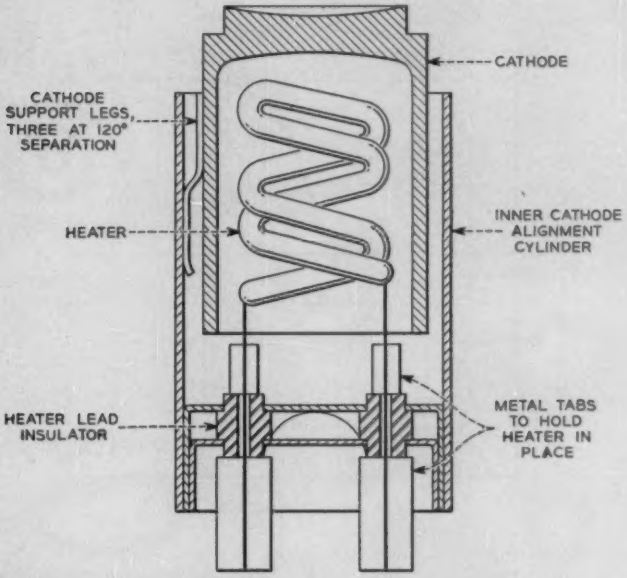


Fig. 9 — The cathode subassembly. In this unit the cathode is connected to the inner alignment cylinder by three legs. These legs are first welded to the cathode and then oven brazed to the alignment cylinder. During the brazing, a jig holds the cathode accurately concentric with this cylinder. The cathode is then coated and the unit is ready for assembly into the gun. The heater power required to raise the cathode to its operating temperature of 720°C is about six watts.

Meaurements of beam interception as a function of magnetic flux density are shown for several beam currents in Fig. 11. These measurements were obtained without any rf input to the TWT. An interesting way of normalizing these data is shown in Fig. 12. Here the magnetic

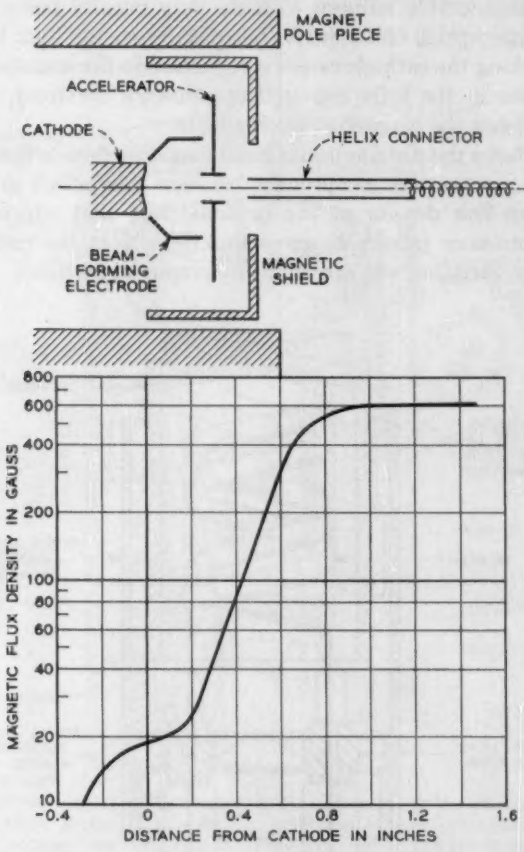


Fig. 10 — Variation in magnetic flux density as a function of distance from the cathode. A schematic representation of the gun electrodes and of the magnetic parts which have been used to control the flux is also shown. All the elements inside the tube are non-magnetic so that the flux density variation is determined entirely by magnetic parts external to the tube envelope. The flux density at the cathode is built up (i.e., the step is put into the curve) by having the magnetic shield end near the cathode. The flux which leaves the shield at this point increases the flux density at the cathode over what it would be if the shield extended well behind the cathode.

flux density has been divided by the Brillouin flux density for a beam entirely filling the helix. This quantity is the minimum flux density which could theoretically be used to focus the beam. This normalization tends to bring all of the curves together. Thus we see that, although the conditions in the MI789 are far from those of ideal Brillouin flow (because of transverse thermal velocities, aberrations in the gun, and magnetic field at the cathode), the concept of the Brillouin flux density still retains meaning, i.e., it appears that the flux density required maintains a fixed ratio to the Brillouin value.

Applying sufficient rf input to the MI789 to drive it into non-linear operation, results in defocusing caused by the high rf fields (both from the helix wave and from space charge) near its output end. Fig. 13 shows how the beam interception for different magnetic flux densities varies as a function of the power output of the TWT. From these curves we see that an output level of five watts can be maintained with about one per cent interception with a flux density of 600 gauss.

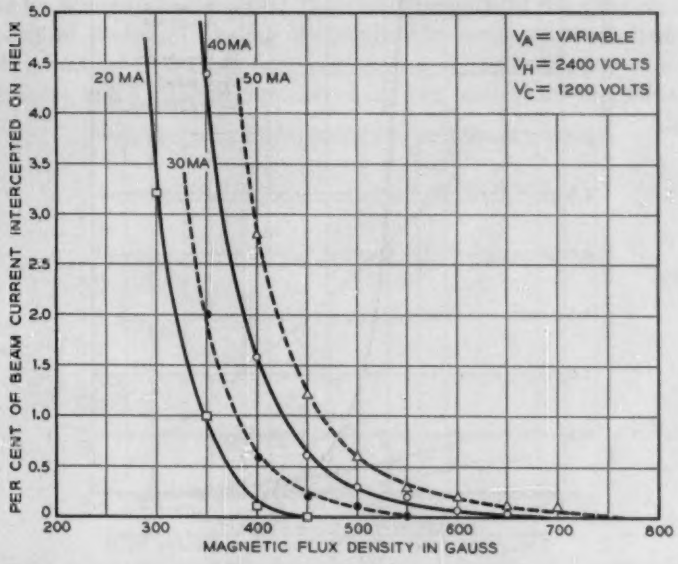


Fig. 11 — Per cent interception on the helix as a function of magnetic flux density. These measurements were taken using a precision solenoid to focus the TWT. The component of field perpendicular to the TWT axis was less than 0.1 per cent of the longitudinal field. During these measurements there was no rf input to the TWT and there was substantially no (<0.1 ma) interception on the accelerator electrode.

3.3 *The Helix*

The MI789 helix assembly is a rigid self-supporting structure composed of three ceramic support rods bonded with glaze to the helix winding. A drawing of the helix assembly is shown in Fig. 14. The support rods are made from Bell Laboratories F-66 steatite ceramic. This material was chosen because of its low rf losses and because these losses do not increase rapidly with temperature. Fig. 15 shows an enlarged photograph of the glaze bonds between the winding and one of the support rods. Attenuation is applied over a length of two inches starting 1½ inches from the input end by spraying the helix assembly with aquadag (carbon in water suspension) and then baking it.

Supporting the winding by glazing it to ceramic support rods has the following advantages:

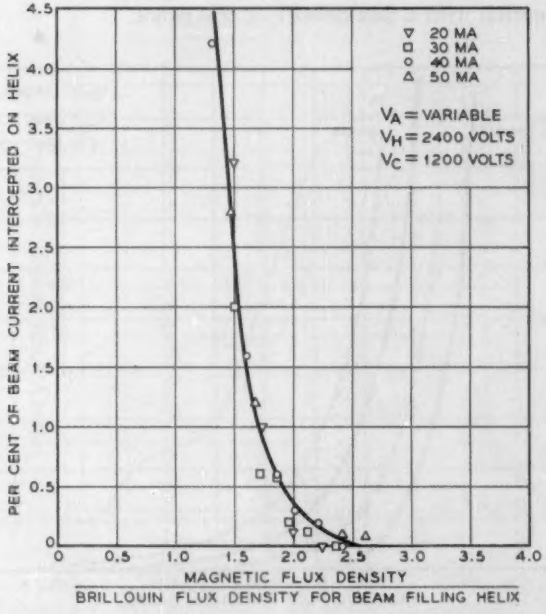


Fig. 12 — The measurements of Fig. 11 normalized in terms of the Brillouin flux density for a beam entirely filling the helix. The fact that the curves tend to come together indicates that the concept of the Brillouin flux density retains some meaning in the M1789. Because of the additional defocusing effects encountered when the M1789 is driven to high output levels, the tube is usually used with about 2.6 times the Brillouin flux density.

(1) The dielectric loading and intrinsic attenuation of the helix are comparatively low because the amount of supporting structure in the rf fields is small.

(2) High loss per unit length in the helix attenuator is made possible. The reason for this will be discussed further below.

(3) The heat dissipation capability of the helix is greatly increased because the glaze provides an intimate thermal contact between winding and support rods. This is illustrated by Fig. 16 which compares the heat dissipation properties of glazed and non-glazed helices.

(4) Mechanical rigidity is realized and therefore the helix can be handled without risk of disturbing the pitch or diameter of the winding.

On the other hand, use of the ceramic rods in the MI789 has a significant disadvantage in that it makes the outside radius of the vacuum envelope large compared to the helix radius, thus making coupled helix matching out of the question. However, since the MI789 is required to match over less than a 10 per cent band, this is not particularly serious.

To obtain reproducibility of performance in the MI789, the helix must be precisely constructed. Together, the pitch of the helix and the amount of dielectric loading determine the synchronous voltage. A pitch variation of ± 1 per cent results in a voltage variation of about ± 50 volts, and a loading variation of ± 1 per cent results in a variation

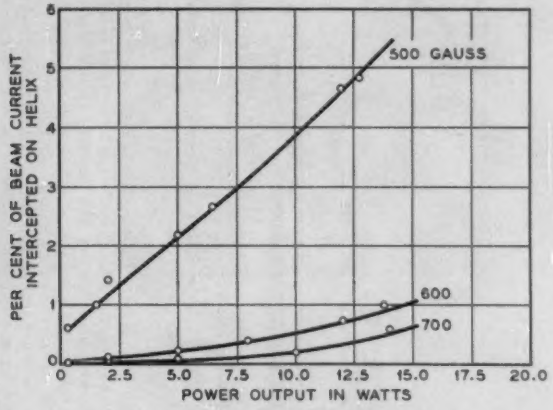


Fig. 13 — Per cent interception on the helix as a function of rf power output. These measurements were made using permanent magnet circuits charged to different field strengths. The magnetic field variation as a function of distance from the cathode was as shown in Fig. 10. The component of magnetic field perpendicular to the tube axis in these circuits was less than 0.2 per cent of the longitudinal field. All measurements were taken with a beam current of 40 ma and with the helix voltage adjusted to maximize the power output.

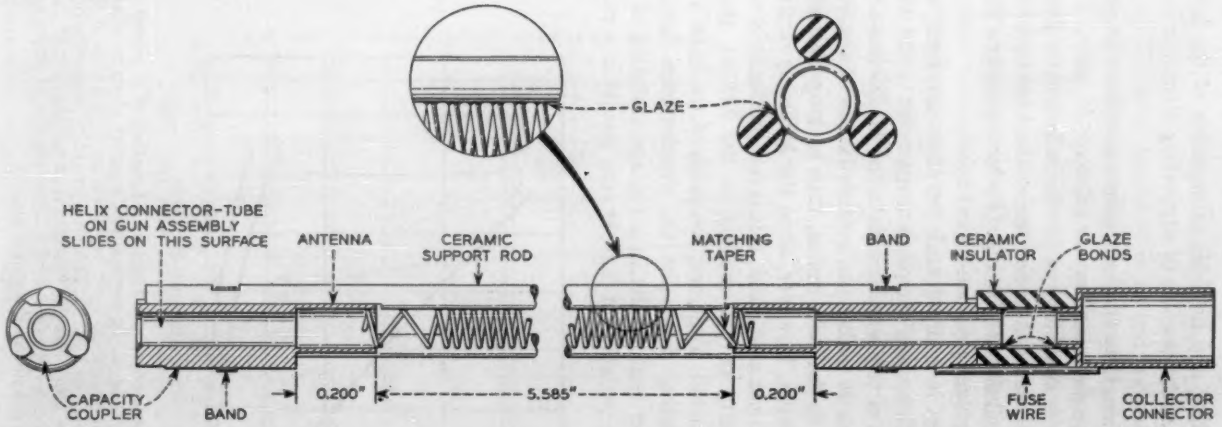


Fig. 14 — Helix assembly. The winding is supported from three F-66 ceramic rods to which it is glazed. On each end of the winding there are two turns with greater than nominal pitch to assist in transferring energy between helix and waveguides. The bands around the assembly are for the purpose of holding the support rods against the capacity couplers. There is no glaze in this region. The relationship of the antenna and capacity coupler to the waveguide circuit is shown in Fig. 5. The collector is later brazed inside of the collector connector cylinder. The fuse wire allows the helix to be heated on the pump station by passing current from the helix pin on the tube base to the collector. After outgassing the helix, the fuse is blown to isolate helix from collector. Before adding the helix attenuation, the rf loss of the helix is 3.6 ± 0.2 db. After adding the attenuation, it is 65 to 80 db. The synchronous voltage is 2200 ± 50 volts.

of about ±25 volts. It is not difficult to hold the average pitch variations to less than ±1 per cent. The loading, however, is a more difficult problem for not only must the dielectric properties of the support rods and of the glaze material be closely controlled, but attention must also be paid to the size and density of the glaze fillets. The gain of the tube is affected by the amount of loss in the helix attenuator. For the particular loss distribution used in the MI789 a variation of ±5 db out of a total attenuation of 70 db results in a gain variation of about ±1 db. The helix attenuator depends to a large extent on a conducting "bridge" between helix turns and therefore the amount of attenuation is sensitive to the size and the surface condition of the glaze fillets. Thus, the glazing process must be in good control in order to minimize variations in both gain and operating voltage. With our present techniques, we are able to hold the voltage for maximum gain to within ±50 volts of the nominal value. The gain is held to ±2 db — about half of the spread we believe to be caused by variations in loss distribution and about half by differences in beam size.

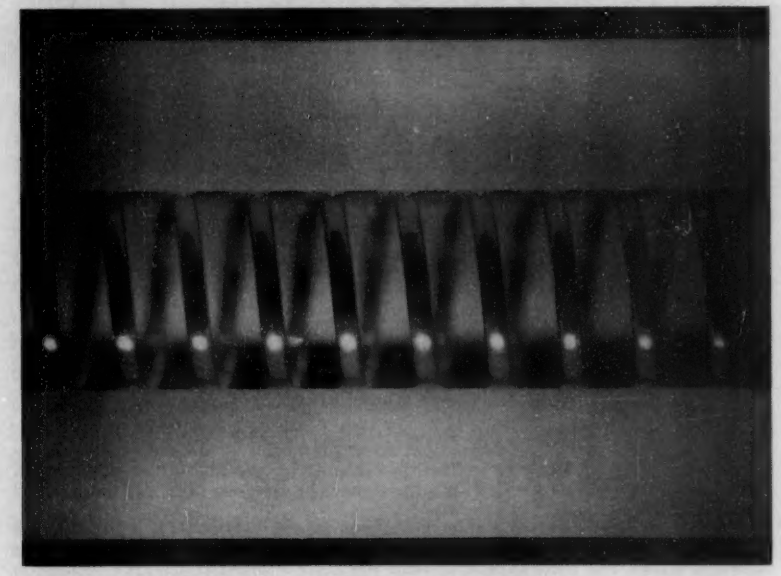


Fig. 15 — Enlarged photograph of part of an M1789 helix. Two of the ceramic support rods can be seen. The other is directly opposite the camera behind the helix and is out of focus. The fillets of glaze which bind the helix to the rods can be seen along the upper rod. This section of helix was free from applied loss.

Helix-to-Waveguide Matching

In the helix-to-waveguide transducer the helix passes through the center of the broad face of the waveguide and energy is coupled between helix and waveguide by an antenna and matching taper. A capacitive coupler on the helix and an rf choke on the waveguide place an effective ground plane at the waveguide end of the antenna. The rf choke also assists in minimizing leakage of rf power. Details of this transducer are shown in Figs. 5 and 14.

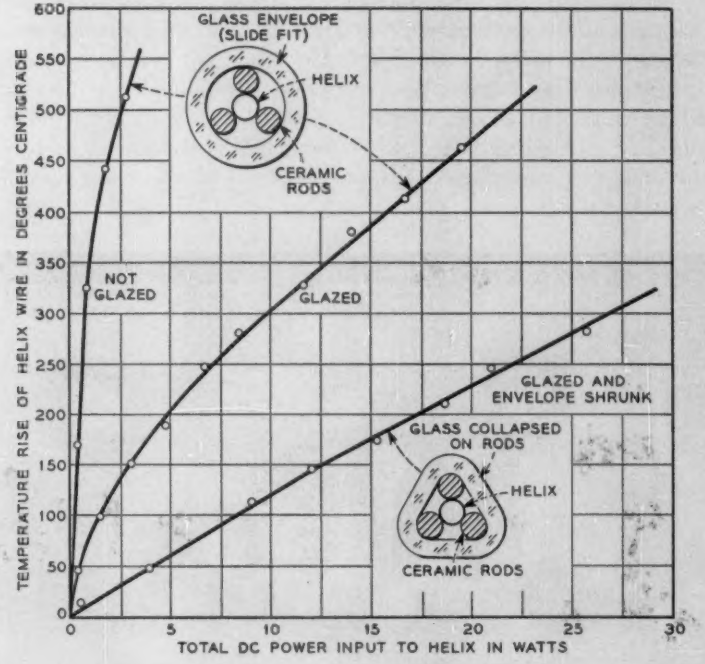


Fig. 16 — Comparison of heat dissipation properties of different helix structures. In this experiment, the helices were heated by passing dc current through them while they were mounted in a vacuum. The temperature was determined from the change in helix wire resistance.

Along with the results for glazed and non-glazed helices in a normal round envelope, this figure shows results on a structure consisting of a glazed helix in an envelope which has been shrunk around the helix support rods. This technique produces a structure which, by virtue of the good thermal contact between the support rods and the envelope, can dissipate more power than the conventional structure. The additional complication of shrinking the envelope is not necessary for the power levels used in the M1789. However, this method could be used if it were necessary to extend the tube's output range to higher power levels.

The dimensions of this transducer were determined empirically. It was found that the antenna length affects mainly the conductive component of the admittance referred to the plane of the helix. The length of the matching taper affects mainly the susceptive component, and the distance from helix to a shorting plunger, which closes off one end of the waveguide, affects both components. If for each tube, the position of the waveguides along the axis of the TWT and the position of the shorting plunger are optimized, the VSWR of the transducers will be less than 1.1 (~26 db return loss) over the entire 500-mc frequency band. With these positions fixed at their best average value, the VSWR will be less than about 1.3 (~18 db return loss).

Internal Reflections

A problem that has required considerable effort has been that of "internal reflections." By this we mean reflections of the rf signal from various points along the helix as contrasted with reflections from helix-to-waveguide transducers. The principal sources of internal reflections are the edge of the helix attenuator and small variations in pitch along the helix. In the MI789 the pitch variations are the main source of difficulty.

The type of performance degradation caused by small internal reflections can be illustrated by the following. Consider a signal incident on the TWT output as a result of a reflection from a radio relay antenna. Except for a small reflection at the transducer, energy incident on the TWT output will be transferred to the helix, propagated back toward the input, and for the most part be absorbed in the helix attenuator. However, if there are reflection points along the helix, reflected signals will be returned to the output having been amplified in the process by the TWT interaction. Because of this amplification, even a small reflection of the backward traveling wave can result in a large reflected signal at the TWT output. In the MI789, these amplified internal reflections are considerably larger than the reflection from the output transducer. They limit the overall output VSWR to about 1.4, whereas the transducer alone has a VSWR of about 1.1.

If there is a long length of waveguide between the TWT and the antenna, the echo signal resulting from a reflection at the antenna and a second reflection at the TWT will vary in phase with respect to the primary signal as frequency is changed. This will cause ripples in both the gain and in the phase delay of the system as functions of frequency. Suppose the VSWR of the antenna is 1.2 and that of the TWT is 1.4 and the two are separated by 100 feet of waveguide. The amplitude of

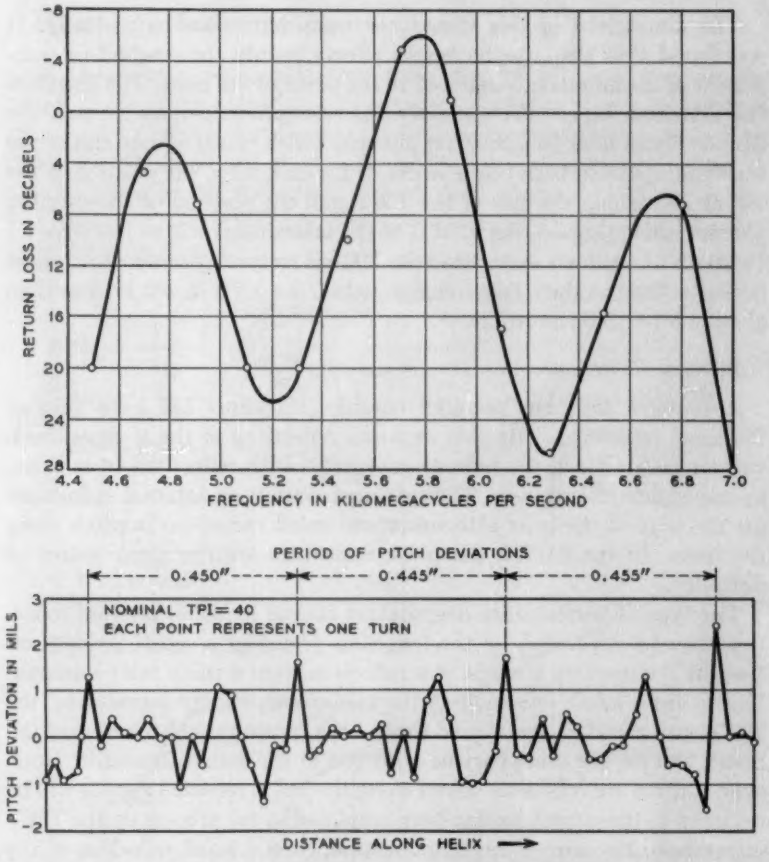


Fig. 17 — Pitch deviations and internal reflections in an early M1789 TWT. The ordinate of the pitch deviation curve is the difference between the measured spacing between helix turns and the nominal value, which for this particular helix was 25 mils. (The tube operated at 1,600 volts.) Each point represents a helix turn. It is seen that the pitch deviations are periodic in nature, repeating about every 0.450 inch.

The internal reflections were measured by matching the TWT with beam off at each individual frequency with a tuner to a VSWR of less than 1.01 (return loss greater than 40 db). The beam was then turned on and the resulting reflection taken as an approximate measure of the internal reflection. There appeared to be no appreciable change in the helix-to-waveguide transducer reflection as a result of turning the beam on. Evidence for this is the fact that when the beam was turned on with the helix voltage adjusted so that the TWT did not amplify, there was little change in the reflection.

The peaks of the internal reflection curve occur at five, six and seven half wavelengths per period of the helix pitch deviations, indicating that the reflections from each period are adding in phase at these frequencies. At the 5,800-mc peak the return loss is positive. This indicates a reflected signal larger than the incident signal. Shorting the TWT output caused the tube to oscillate at this frequency.

the gain fluctuations will be about 0.25 db, the amplitude of the phase fluctuations will be about 0.9 degree and the periodicity of the fluctuations will be about six mc. This effect may be eliminated by using an isolator between the TWT and the antenna to eliminate the echo signal.

In addition to echo signals that occur between the TWT and the antenna there are echoes which occur wholly within the TWT as a result of a reflection of the signal from the output transducer and a second reflection from some point along the helix. Thus even if a TWT is operating into a matched load it may have ripples in gain or phase characteristics. These ripples may be controlled by minimizing the internal reflections. In the MI789 they are less than ± 0.1 db in gain and one-half degree in phase. Their periodicity is about 100 mc.

In addition to causing transmission distortions, internal reflections can seriously reduce the margin of a TWT against oscillation. Outside of the frequency band of interest, the helix-to-waveguide transducer may be a poor match or the TWT may even be operating into a short circuit in the form of a reflection type bandpass filter. At such frequencies, the internal reflections must not be large enough so that an echo between transducer or filter and an internal reflection point will see any net gain, or else the TWT will oscillate.

With many types of helix winding equipment, variations in helix pitch are periodic in nature. This causes the helix to exhibit a filter-like behavior with respect to internal reflections. At frequencies at which the period of the pitch variations is an integral number of half-wave lengths, the resultant reflections from each individual period will add in phase, thereby causing the helix to be strongly reflecting at these frequencies. This effect can perhaps best be illustrated by considering some results obtained in an early stage of the MI789 development. Fig. 17 shows measurements of the spacing between turns of an early helix. Also shown is the return loss as a function of frequency that a signal incident on the output of an operating TWT would see as a result of internal reflections alone. Helix-to-waveguide transducer reflections were eliminated with waveguide tuners during this experiment. The deviations in helix pitch from nominal are rather large and are markedly periodic in nature. The resulting internal reflections show strong peaks at frequencies corresponding to five, six and seven half-wavelengths per period of the pitch deviations.

In the present M1789 this situation has been considerably improved by increased precision in helix winding and by insuring that the remaining periodicity does not produce a major reflection peak in the band. Fig. 18 shows pitch measurements and internal reflections for a recently constructed tube.

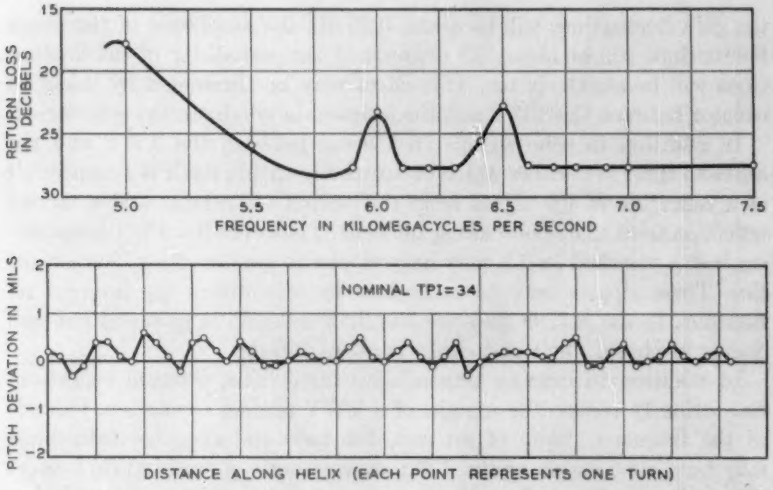


Fig. 18 — Pitch deviations and internal reflections in a recent M1789 TWT. By precise helix winding techniques the pitch deviations have been reduced by a factor of about 10 over those occurring in early tubes. The resulting internal reflections have been improved by about 25 db although there is still a residual periodicity remaining.

For return losses greater than about 25 db, we begin to see internal reflections originating from the edge of the helix attenuator. At these values of return loss, the measurements also begin to be in appreciable error as a result of the residual transducer reflections.

Helix Attenuator

Attenuation is applied to the helix by spraying aquadag directly on the helix assembly and then baking it. The result is a deposit of carbon on the ceramic rods and on the glaze fillets. The attenuation is held between 65 and 80 db and is distributed as shown in Fig. 19. Evidently most of the loss is caused by a conducting bridge which is built up between helix turns. This was indicated by one experiment in which we cleaned the deposit off the rods of a helix by rubbing them with emery paper. Only the carbon directly between helix turns then remained. This decreased the total attenuation by less than 20 per cent. Having the helix glazed to the support rods is apparently necessary in order to get good contact between the winding and the carbon "bridge." We have been able to obtain about four times as much loss per unit length with glazed helices as with non-glazed ones. Using our method of applying attenuation we can add in excess of 80 db/inch to a glazed helix. The ability to obtain such high rates of attenuation allows us to concentrate the loss along the helix thereby minimizing the TWT length.

The machine used for spraying aquadag on the helix is shown in Fig.

20. A glass cylinder and photocell arrangement is used to monitor the amount of carbon deposited. In this manner the attenuation added is made independent of both the aquadag mixture and the nozzle setting of the spray gun. This machine has been checked alone by using it to spray glass slides which are then made into attenuator vanes. Over a two-year period we have found that a given light transmission through the monitor slide results in the same vane attenuation within ±2 db out of 40 db.

After a helix has been sprayed, it is vacuum fired at 800°C for thirty minutes and then the loss is measured. About 60 per cent of the helices fall within the desired range of 65–80 db. The principal cause of the differences in attenuation is believed to be variation in the condition of the glaze fillets. Helices not meeting specifications are sprayed and fired a second time (after cleaning off excess aquadag if necessary). This second treatment, brings the attenuation of almost all helices to within the desired range.

3.4 *The Collector*

It is desirable to operate the collector at the lowest possible voltage in order to minimize the dc power input to the TWT. This increases the overall efficiency and simplifies the collector cooling problem. On the

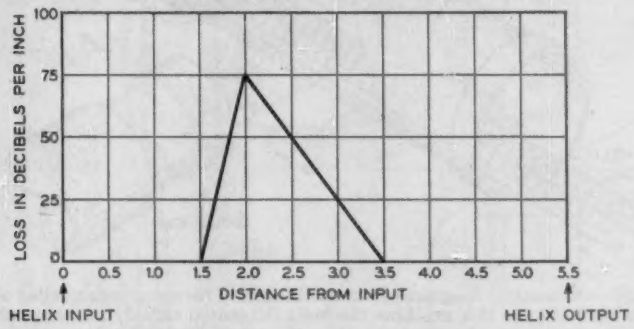


Fig. 19 — Distribution of helix attenuation. The attenuation pattern has a gradually slanting edge facing the output to provide a smooth transition into the loss for any signals traveling backwards toward the input. Reflections of these signals must be very small since the reflected signals will be amplified in the process of returning toward the output. Cold measurements (i.e., measurements on the helix without electron beam) made by moving a sliding termination inside the helix, indicate that the return loss from the attenuator output is better than 45 db, the limiting sensitivity of our measurement. The input side of the helix attenuator is also tapered to minimize reflections but this taper is much sharper than that on the output side because there is comparatively little gain between input and attenuator. Cold measurements with a sliding termination showed a return loss for this taper of about 40 db. (Surprisingly, even a sharp edge produces a reflection with a return loss of almost 30 db.)

other hand, if there is appreciable potential difference between helix and collector, we must insure that few secondary or reflected electrons return from the collector to bombard the helix and accelerator, or else we may overheat these electrodes. Fig. 21 shows a drawing of the collector used in the M1789. It takes the form of a long hollow cylinder shielded from the magnetic field. Inside of the collector the beam is allowed to gradually diverge and the electrons strike the walls at a grazing angle. This design reduces secondary electrons returned from the collector to almost negligible proportions.

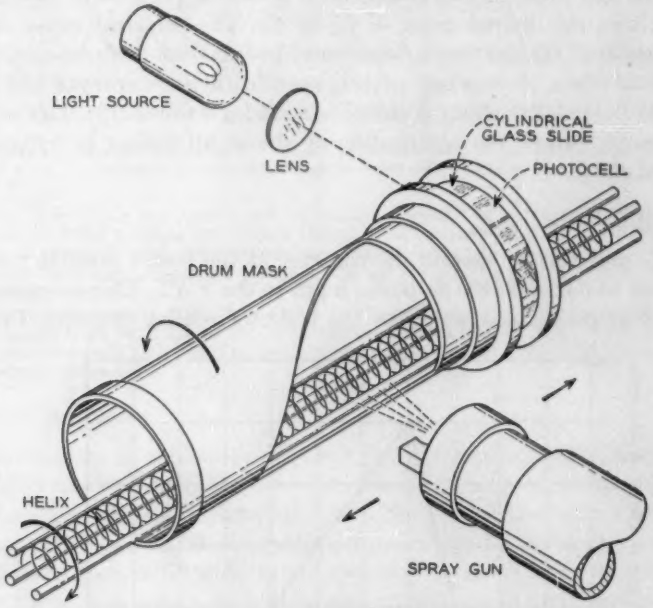


Fig. 20 — Schematic diagram of the machine used for spraying aquadag attenuation on the helix. In this machine the helix is rotated rapidly to insure uniform exposure to the spray. At the same time the masking drum rotates at a slower speed and the spray gun traverses back and forth along the masking drum. The drum therefore acts as a revolving shutter between the helix and the spray gun and its degree of opening serves to control the amount of aquadag reaching the helix. From a knowledge of the rate of attenuation increase as a function of the amount of carbon deposited (empirically determined) the shape of the drum opening can be calculated so as to give any desired attenuation pattern.

The spray gun also passes over a glass cylinder at one end of the masking drum so that it receives a sample of the aquadag spray. A photocell is used to monitor light transmitted through the cylinder. Before starting to spray, the glass is cleaned and the photocell reading is taken as 100 per cent light transmission. The helix is then sprayed until the light transmission has decreased to the proper value. The photoelectric monitoring technique makes the attenuation added insensitive to the aquadag composition and to the spray gun nozzle opening.

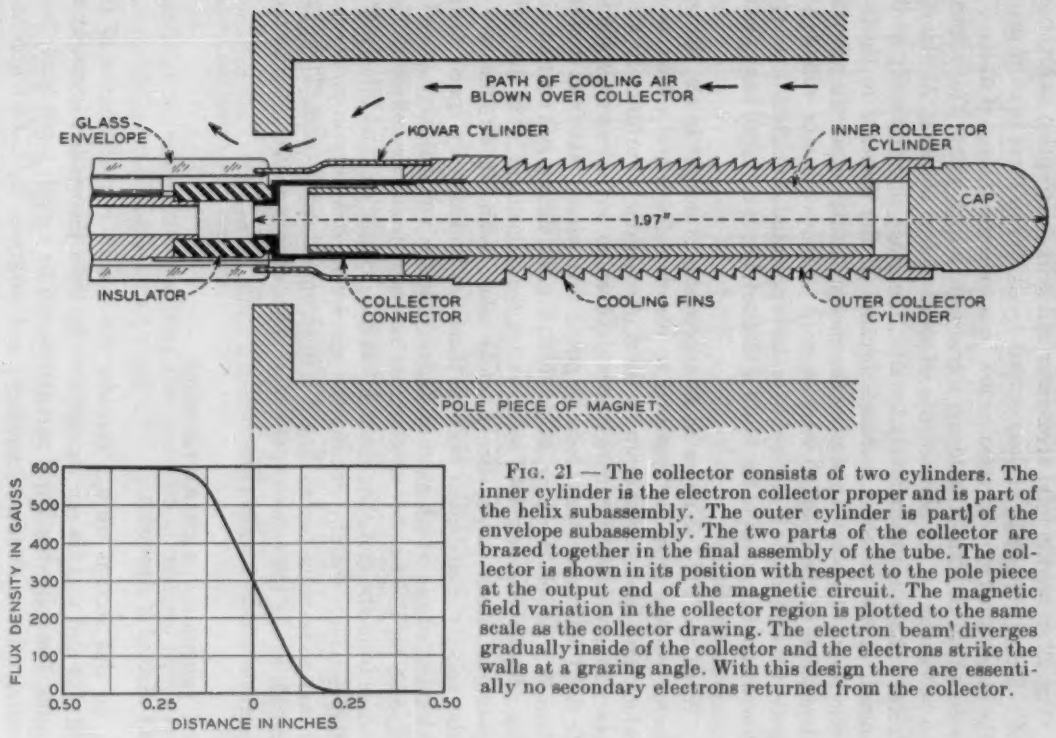


FIG. 21 — The collector consists of two cylinders. The inner cylinder is the electron collector proper and is part of the helix subassembly. The outer cylinder is part of the envelope subassembly. The two parts of the collector are brazed together in the final assembly of the tube. The collector is shown in its position with respect to the pole piece at the output end of the magnetic circuit. The magnetic field variation in the collector region is plotted to the same scale as the collector drawing. The electron beam diverges gradually inside of the collector and the electrons strike the walls at a grazing angle. With this design there are essentially no secondary electrons returned from the collector.

Fig. 22 shows the total accelerator and helix interception as functions of collector voltage at various output levels. When there is no rf drive, the intercepted current remains low to a collector voltage of about 200 volts at which point it suddenly increases to a high value. This appears to be caused by the phenomenon of space charge blocking. As the collector voltage is progressively lowered, the space charge density at the mouth of the collector increases because of the decrease in electron velocity at this point. Increasing the charge density causes the potential depression in the beam to increase until at some collector voltage the potential on the axis is reduced to cathode potential. At collector voltages lower than this, some of the beam is blocked, i.e., it is turned back by the space charge fields.

When the TWT is operated at appreciable rf output levels, the collector voltage must be increased to permit collection of all electrons which have been slowed down by the rf interaction. Unfortunately, some electrons are slowed far more than is the average, so that we must supply to the TWT several times more dc power than we can take from it in the form of rf power. However, as seen from Fig. 22, there is still an appreciable advantage to be gained by operating the collector at lower than helix potential. These curves should not be taken as an accurate measure of the velocity distribution because there are undoubtedly space charge blocking effects which even at higher collector voltages have some influence on the number of electrons returned from the collector. This arises from the fact that the rf interaction causes an axial bunching of the electrons, thereby causing the space charge density in an electron bunch to be much higher than it is in an unmodulated beam. Thus, as a bunch enters the collector, the local space charge density may be high enough to return some electrons.

IV. PERFORMANCE CHARACTERISTICS

4.1 *Method of Approach*

In this section we will consider the overall rf performance of the M1789 and make some comparisons between theory and observed results. The following TWT parameters can be varied: input level; helix voltage; beam current; frequency; and magnetic field. Our approach here will be to first consider the operation of the tube under what might be called nominal conditions. This will be followed by a discussion of the variations in low-level gain and in maximum output over an extended range of beam current, frequency, and magnetic field. By this procedure we are able to obtain a description of tube performance without presentation of a formidable number of curves. Two topics, noise and inter-

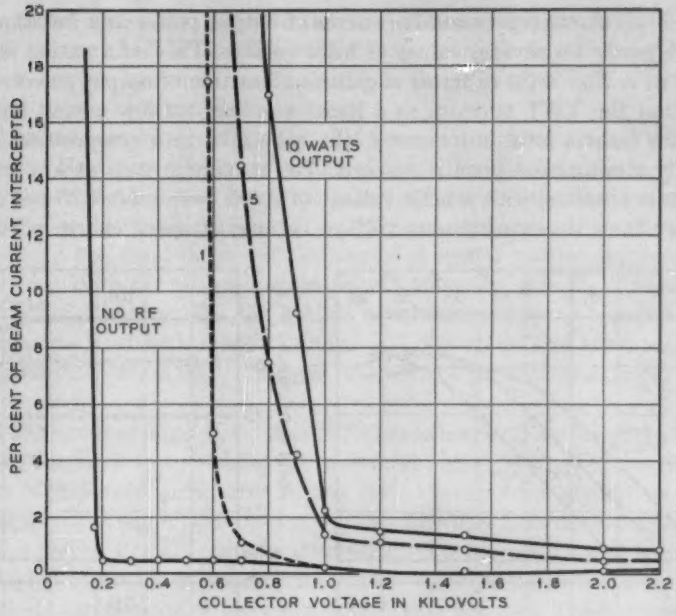


Fig. 22 — Intercepted current as a function of collector voltage with helix and accelerator voltages held constant at their nominal values. Below the knee of the curves about three quarters of the total intercepted current goes to the helix and about one quarter is focused all the way back to the accelerator. Curves are shown for no rf input and for output levels of 1, 5, and 10 watts. With no input, the lowest permissible collector voltage is determined by the phenomenon of space charge blocking. With rf input, it is determined mainly by the velocity spread of the electrons. In all cases it was found that the alignment of the TWT with respect to the magnetic circuit becomes more critical as the knee of the curve is approached. For this reason the M1789 is usually operated with a collector voltage about 200 volts above the knee.

modulation, will be divorced from the discussion as outlined above and treated separately in Sections 4.4 and 4.5.

4.2 *Operation Under Nominal Conditions*

Basic Characteristics

By nominal conditions for the M1789 we mean the following:

frequency	6175 mc (band center)
beam current	40 ma
magnetic flux density	600 gauss
collector voltage	1200 volts

Fig. 23(a) shows representative curves of output power as a function of input power for several values of helix voltage. This information is replotted in Fig. 23(b) in terms of gain as a function of output power. We see that the TWT operates as a linear amplifier for low output levels. As the output level is increased, the tube goes into compression and finally a saturation level is reached. The maximum gain at low input levels is obtained with a helix voltage of 2,400 volts (about 10 per cent higher than the synchronous voltage because of space charge effects).

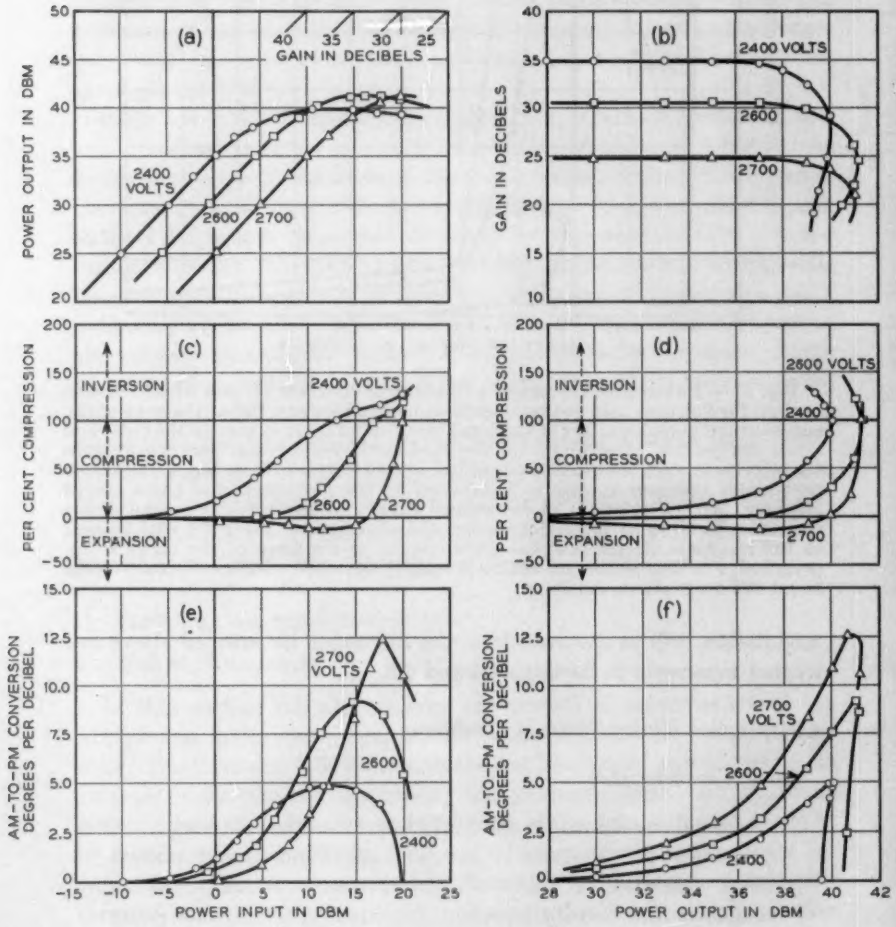


Fig. 23 — See opposite page for caption

The maximum output at saturation is obtained at a higher helix voltage as is common in TWT's. The helix voltage also affects the shape of the input-output curves — linear operation being maintained to higher output levels at higher helix voltages.

As a measure of the efficiency of electronic interaction in a TWT, we use an "electronic efficiency" which is defined as the ratio of the rf output power to the beam power (product of helix voltage and beam current). The "over-all efficiency" we define as the ratio of the rf output power to the total dc power (exclusive of heater power) delivered to the tube. With the collector operated at 1,200 volts, it is about twice the electronic efficiency. For the M1789, maximum efficiency occurs at the saturation level with a helix voltage of 2,600 volts. The electronic and over-all efficiencies there are equal to about 14 per cent and 28 per cent, respectively.

The curves of Figs. 23(a) and (b) were taken with sufficient time allowed for the tube to stabilize at each power level. If the TWT is driven to a high output level after having been operated for several minutes with no input signal, the output will be somewhat greater than is shown in the curves. It will gradually decrease until it reaches a stable level in a period of about two minutes. This "fade" is caused by an increase in the intrinsic attenuation of the helix near the output end. The increase is a result of heating from rf power dissipation. At maximum output the fade is about 0.6 db (about 15 per cent decrease in output power). At the five-watt output level the fade is about 0.1 db (about 2 per cent

Fig. 33 — See opposite page

(a) Output power as a function of input power. Both ordinate and abscissa are in dbm (db with respect to a reference level of one milliwatt). A straight line at 45° represents a constant gain. A gain scale is included along the top of the figure. For this tube a helix voltage of 2,400 volts gives maximum gain at low signal levels and a voltage of about 2,600 gives maximum output at saturation.

(b) Gain as a function of output power. This is an alternate way of presenting the information shown in (a).

(c) Compression as a function of input power. Three regions are shown in the figure. The "compression" region is that in which there is less than one db change in output level for a db change in input level. The "expansion" region is that in which there is more than one db change in output level for a db change in input level. The "inversion" region is that in which the output level decreases when the input level increases (or vice versa). It occurs for input levels greater than that necessary to drive the TWT to saturation. In this region the change in output is of opposite sign to the change in input. Using the definition in the text this gives rise to compression values in excess of 100 per cent.

(d) Compression as a function of output power.

(e) Conversion of amplitude modulation to phase modulation as a function of input power. This conversion arises because the electrical length of the TWT is a function of the input level. The effect can cause rather serious difficulties in certain types of low index FM systems.

(f) Conversion of amplitude modulation to phase modulation as a function of output power.

decrease in output power). We will present some additional data on this effect in Section 4.3.

Distortion of the Modulation Envelope

The curves of Figs. 23(a) and (b) tell what happens when a single frequency carrier signal is passed through the TWT. In addition we would like to know the effect on modulation which may be present on the signal. In particular, it is desirable to know the compression of the envelope of an AM signal and the amount of phase modulation generated in the output signal as a result of amplitude modulation of the input signal, (an effect commonly known as AM-to-PM conversion). As a measure of compression of an AM signal the quantity per cent compression will be used. This is defined as

$$\% \text{ Compression} = \left[1 - \frac{\Delta V_0/V_0}{\Delta V_i/V_i}\right] 100$$

where V_0 is the voltage of the output wave, V_i is the voltage of the input wave, and ΔV_0 is the change in output voltage for a small change ΔV_i in the input voltage. When $\Delta V/V$ is small it can be expressed in db as $8.68\ \Delta V/V = \Delta V/V$ in db. From this it follows that

$$\% \text{ Compression} = \left[1 - \frac{\Delta P_0}{\Delta P_i}\right\} \text{ in } db \Bigg] \ 100$$

where ΔP_0 is the change in output power for a change ΔP_i in input power, and the two powers are measured on a db scale. When the per cent compression is zero the TWT is operating as a linear amplifier; when it is 100 per cent the TWT is operating as a limiter.

From the above expression it may appear that the per cent compression could be determined directly from the slopes of the input-output curves. This would be the case were it not for fading effects. Since there is fading, however, the slope for rapid input level changes is different at high levels from the slope of the static curves. Thus it is necessary to determine compression from the resulting effect on an AM signal.

The electrical length of a TWT operated in the non-linear region is to some extent dependent on the input level. Therefore, an AM signal applied to the input of the TWT will produce phase modulation (PM) of the output signal. This effect may be of particular concern when a TWT operating at high output levels is used to amplify a low-index FM signal. If such a signal contains residual amplitude modulation, the TWT generates phase modulation with phase deviation proportional to the input amplitude variation. Under certain circumstances this can cause

severe interference with the signal being transmitted. We will discuss a particular example after consideration of the compression and AM-to-PM conversion characteristics of the M1789.

As in the case of compression, we must measure AM-to-PM conversion dynamically. This is necessary because point-by-point measurements of the shift in output phase as input level is changed include a component of phase shift caused by changes in temperature of the ceramic support rods and a consequent change in their dielectric constant. However, this thermal effect does not follow AM rates of interest and therefore does not produce AM-to-PM conversion.

Fig. 24 shows a simplified block diagram of the test set used to measure compression and AM-to-PM conversion. This equipment amplitude modulates the input signal to the TWT under test by a known amount and detects the AM in the output signal with a crystal monitor and the PM with a phase bridge. A more complete discussion of this measurement is given by Augustine and Slocum.[3]

Compression is given as a function of power input in Fig. 23(c) and as a function of power output in Fig. 23(d). We see that compression sets in more suddenly at higher helix voltages. Above about 2,500 volts

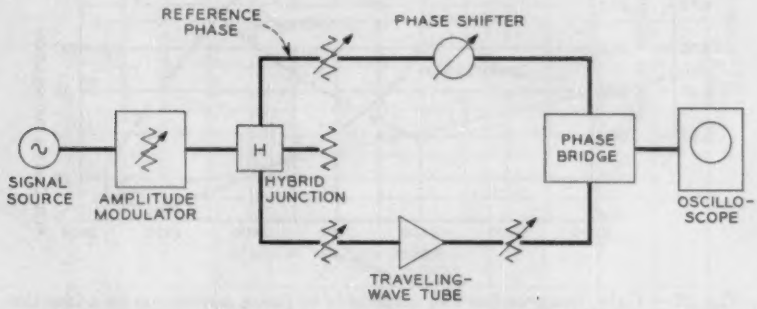


Fig. 24 — Simplified block diagram of test set used to measure compression and conversion of amplitude to phase modulation. A ferrite modulator introduces one db of 60 cps amplitude modulation into the test signal. The 60 cps rate is much higher than that which can be followed by thermal changes in the TWT. Half of the modulated signal serves as input to the TWT under test and half serves as a reference phase for a phase detector. The signals at the phase detector input are maintained equal and at constant level and nominally in phase quadrature. The detector is essentially a bridge circuit, the output of which is a dc voltage proportional to the phase difference of the two inputs. When operated with inputs in quadrature it is not sensitive to amplitude changes of as much as two db in either or both inputs. Phase modulation introduced by the amplitude modulator appears at both inputs and thus does not produce an indication. The output of the detector is therefore a direct measure of the phase modulation created in the TWT. Compression is determined by comparing the percentage amplitude modulation at the input and output crystal monitors.

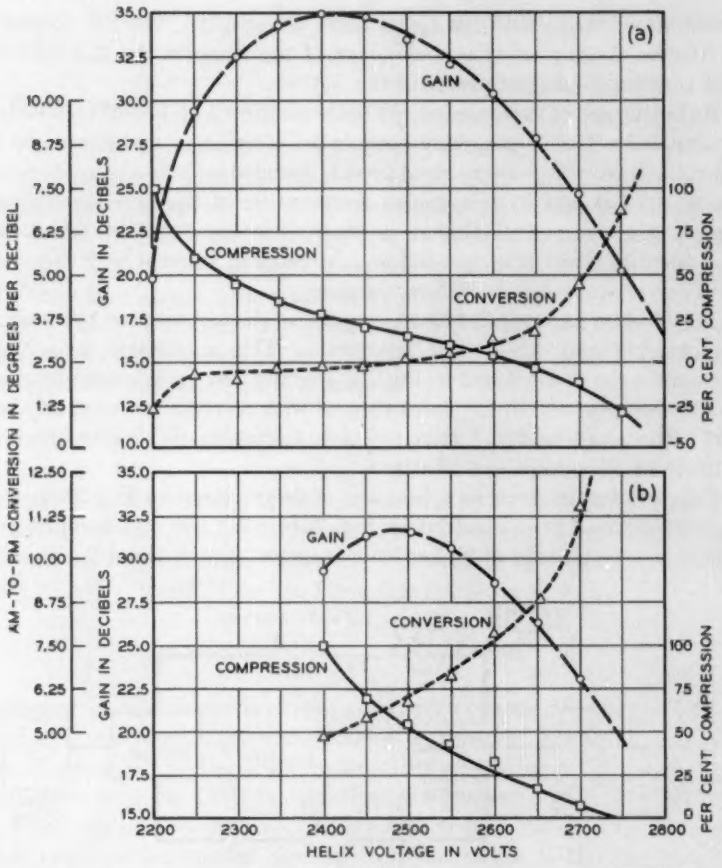


Fig. 25 — Gain, compression and amplitude to phase conversion as a function of helix voltage with the output power maintained constant at a level of five watts (a) and ten watts (b).

there is expansion for some values of power input. Figs. 23 (e) and 23(f) give the AM-to-PM conversion, as functions of input and output power respectively. These data indicate that the conversion is very much less if the tube is operated at lower helix voltages. For example, the conversion at the saturation level of the 2,700-volt curve is about 2½ times that for the 2,400-volt curve.

A final method of plotting gain, compression, and AM-to-PM conversion data is shown in Fig. 25. The abcissa here is the helix voltage.

For these measurements power output was held constant by adjusting input level at each voltage. The figure shows that as helix voltage is increased, the compression decreases but the AM-to-PM conversion increases. The choice of a helix voltage at which to operate the tube must therefore represent a compromise between these quantities.

Phase Modulation Sensitivity

The equipment of Fig. 24 was also used to measure the phase modulation sensitivity of various electrodes by omitting the amplitude modula-

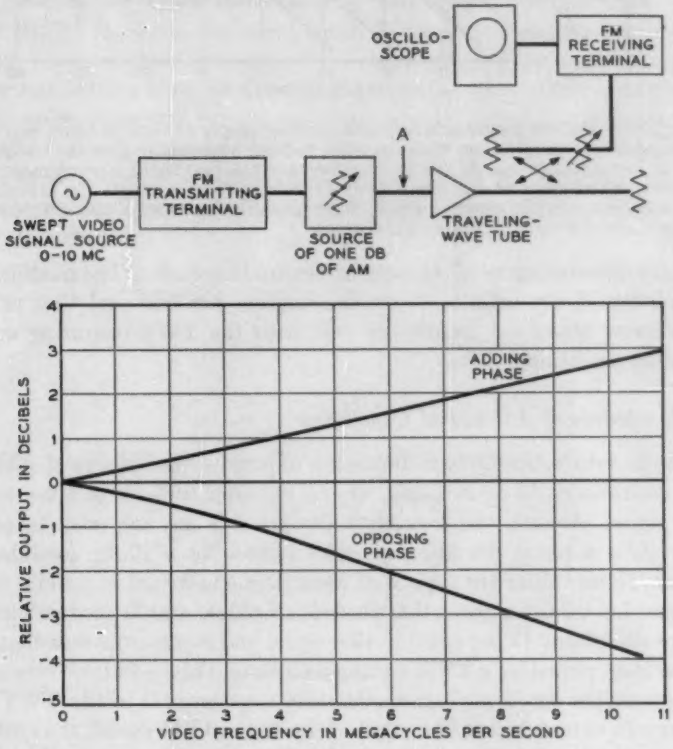


Fig. 26 — Example of frequency response shaping caused by AM-to-PM conversion. This figure shows the calculated frequency response viewed between FM terminals for the system shown in the block diagram. Curves are given for the case in which the phase modulation generated in the TWT both adds to and subtracts from that of the transmitted signal. Inclusion of a limiter at point A would result in a flat frequency response.

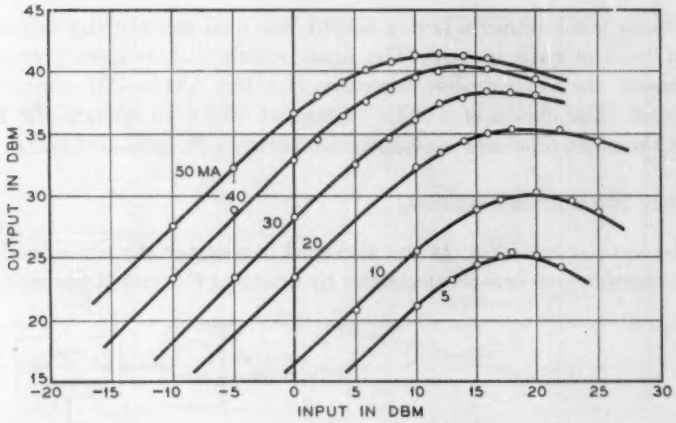


Fig. 27 — Output power as a function of input power at various beam currents. These curves were all taken with the helix voltage adjusted to give the maximum gain at low signal levels. At low beam currents (<20 ma) there is insufficient gain between the attenuator and the output so that at these currents the attenuator section is limiting the power output. This accounts for some of the difference in shape of the curves near maximum output.

tor and introducing small changes in electrode voltages. The modulation sensitivity of the helix is about two degrees per volt and that of the accelerator about 0.1 degree per volt with the TWT operating under nominal conditions.

Significance of AM-to-PM Conversion

Let us return briefly to a discussion of some consequences of AM-to-PM conversion. As an example, we will consider the case of a low-index FM signal. Assume the frequency deviation is ±5 mc peak to peak. This gives a phase deviation of ±0.5 radian for a 10 mc modulating signal. These values are typical of what might be found in a radio relay system. Let us also assume that there is a residual amplitude modulation of one db (about 13 per cent) in this signal and suppose further that the signal is amplified by a TWT having a value of AM-to-PM conversion of 10 degrees per db. The phase modulation thus created in the TWT can either add to or subtract from that of the original FM signal, thus changing its modulation index. At low modulation signal frequencies the phase deviation of the FM signal will be large compared to that of the PM interference and the interference will be of little consequence. At high modulation signal frequencies the phase deviation of the original FM and of the interfering PM signals will be comparable and the interference

can considerably change the net phase deviation of the overall signal. For the example we are considering the frequency responses in Fig. 26 show what would be seen at the output FM terminal. Curves are given both for the PM interference adding to and subtracting from the original FM signal. We see that a gain-frequency slope of about 4 db over 10 mc is introduced by AM-to-PM conversion. To prevent such an effect, a limiter should be used prior to the TWT in applications of this nature so as to remove the offending AM from the input signal.

The fact that compression and amplitude-to-phase conversion vary with input level means that in addition to the first order distortion just described, higher order distortions of the modulation envelope will occur. If, for example, the input signal is amplitude modulated at frequency f_1, the output modulation envelope will contain amplitude and phase modulation both at f_1 and at harmonics of f_1. The amount of higher order distortion can be estimated by expanding the compression and amplitude-to-phase conversion curves as a function of power input in a Taylor series about the operating point. Such an expansion shows that the greater the slope of these curves the greater will be the higher order distortions.

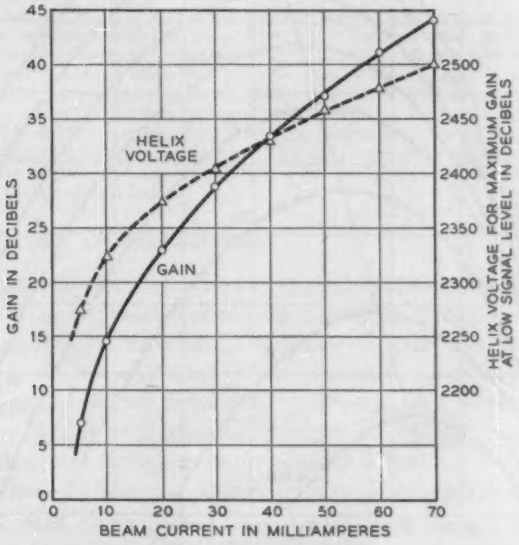


Fig. 28 — Low-level gain as a function of beam current. The helix voltage was adjusted for maximum gain at each current.

Reproducibility

The curves presented in this section are all for the same tube, one which is representative of a group of 50 which were built at the conclusion of the M1789 development program. The tubes in this group had characteristics falling within the following ranges. The numbers represent the range containing 90 per cent of the tubes tested.

Accelerator Voltage for 40 ma	2,500–2,700
Helix Voltage for maximum low-level gain	2,350–2,450
Low-level gain	33–37 db
Gain at 5 watts output	31–35 db
Maximum power output	40.5–42 dbm (11.2–15.8 watts)

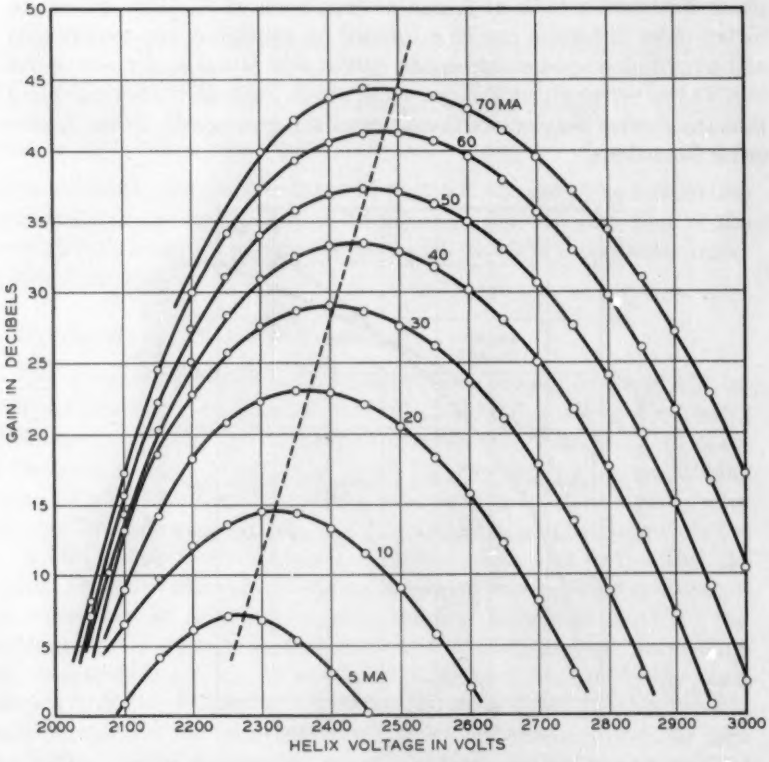


Fig. 29 — Low-level gain as a function of helix voltage for various beam currents. The dotted line represents the locus of the maxima of the curves.

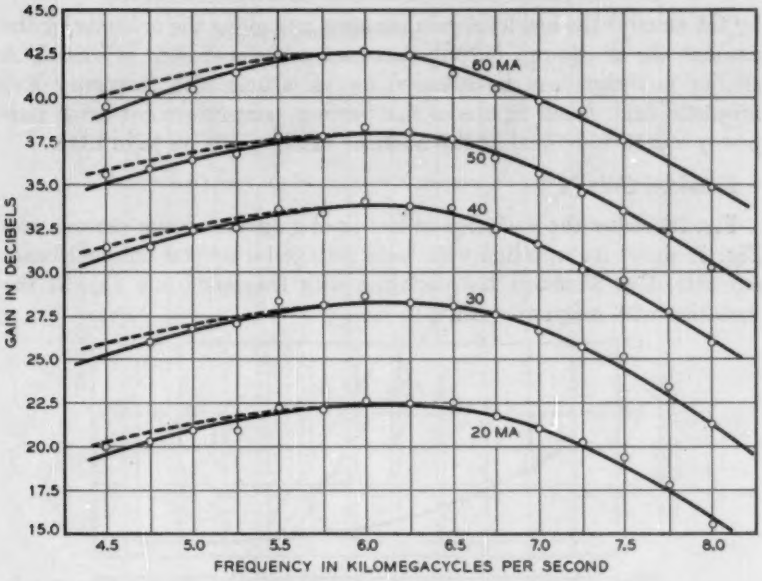


Fig. 30 — Low-level gain and helix voltage for maximum gain as functions of frequency for several beam currents. The TWT was matched to the waveguide (with tuners where necessary outside of the 5,925 to 6,425-mc range) at each frequency. The solid curves show the gain-frequency characteristic with the helix voltage adjusted for maximum gain at 6,000 mc for each beam current and then held constant as frequency was changed. Experimental points correspond to this condition. The dotted curves show how the characteristics change when helix voltage is optimized at each frequency. The optimum helix voltage increases by about 100 volts in going from 6,000 down to 4,500 mc because of slight dispersion in the phase velocity of the helix.

4.3 *Operation Over an Extended Range*

We now turn to a consideration of typical M1789 characteristics over an extended range of beam current, frequency, and magnetic field.* We shall concentrate on two items, the low-level gain and the maximum power output. From variations in these quantities the complete compression curves can be roughly deduced. This situation is illustrated in Fig. 27 which shows output as a function of input at different beam currents. While the shapes of these curves are slightly different, for the most part they can be derived from the 40-ma curve by shifting it along the abcissa

* The characteristics of the tube used for the low-level gain measurements in this Section were slightly different from those of the tube used for the maximum output measurements and both were slightly different from those of the tube used for the measurements of Section 4.2. All tubes, however, had characteristics falling within the ranges listed above.

by the amount the low-level gain changes, and along the ordinate by the amount the maximum output changes as beam current is varied. A similar procedure can be followed for variations with frequency and magnetic field. In all figures in this Section, parameters not being purposely varied were held at the nominal values given on page 1315.

Low-level Gain

Fig. 28 shows the variation in low-level gain with beam current and Fig. 29 shows its variation with helix voltage for several different beam currents. Fig. 30 shows the variation with frequency and Fig. 31 the variation with magnetic field.

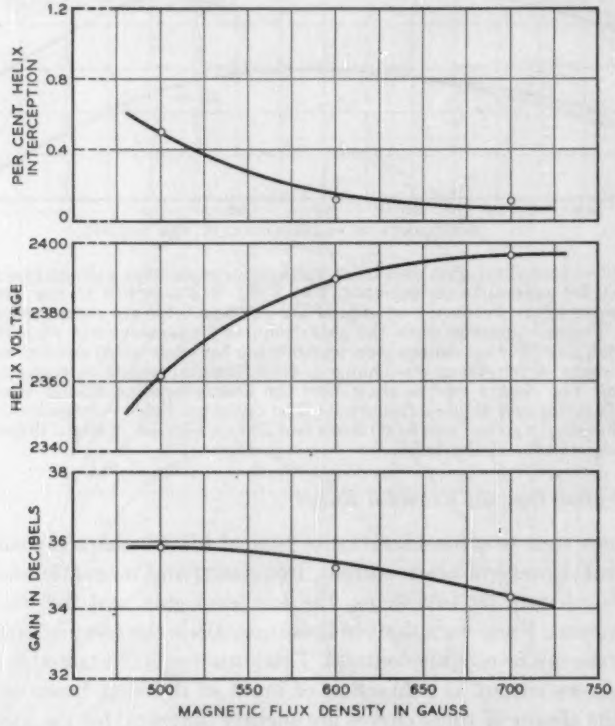


Fig. 31 — Low-level gain, helix voltage for maximum gain and helix interception at low signal level as functions of magnetic flux density. These measurements were made using different strength permanent magnet circuits. The gain varies with magnetic flux density mainly as a result of its effect on beam size and therefore on the degree of coupling between electron stream and helix. The helix voltage varies because of the effect of beam size on QC and therefore on the ratio of the optimum gain voltage to the helix synchronous voltage.

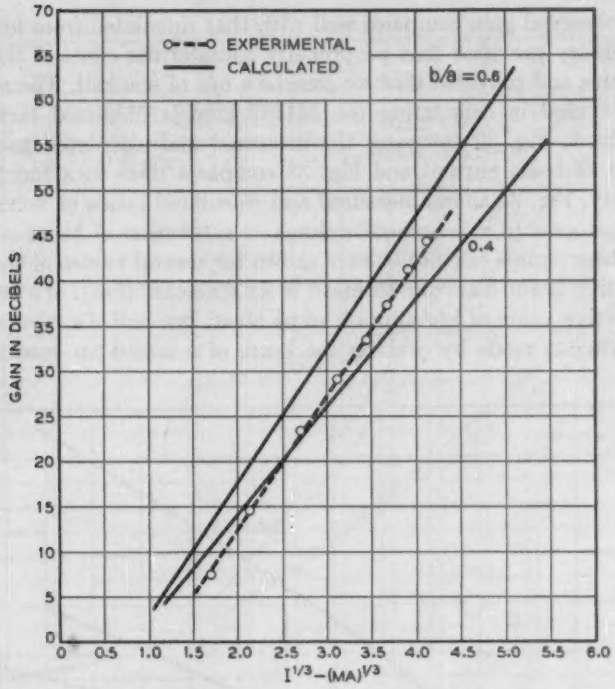


Fig. 32 — Measured and calculated low-level gain as a function of the one-third power of beam current. The parameter b/a is the ratio of effective beam diameter to mean helix diameter.

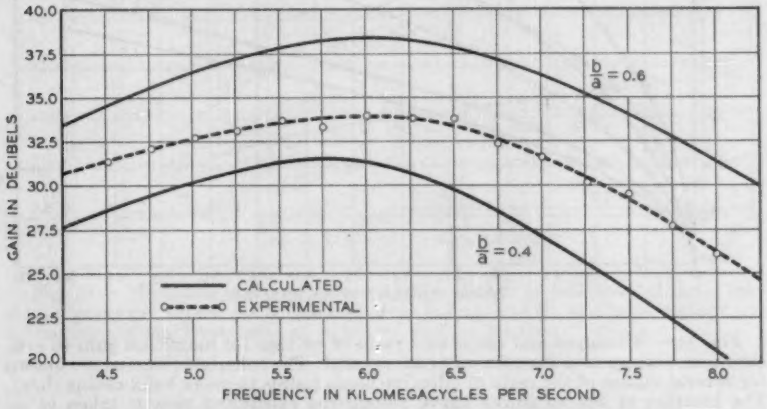


Fig. 33 — Measured and calculated frequency response for a current of 40 ma.

The observed gain compares well with that calculated from low-level TWT theory provided that we properly consider the effect of the helix attenuator and provided that we assume a b/a of one-half. The method we have used in calculating the M1789 gain is discussed further in Appendix I. Fig. 32 compares the measured and calculated gain as a function of beam current and Fig. 33 compares them as a function of frequency. Fig. 34 shows measured and calculated ratios of voltage for maximum gain to synchronous voltage as a function of beam current. In all these figures calculations are shown for several values of the ratio of effective beam diameter to mean helix diameter (b/a). We see that the effective value of b/a appears to be about one-half. On the basis of measurements made by probing the beam of a scaled up version of a

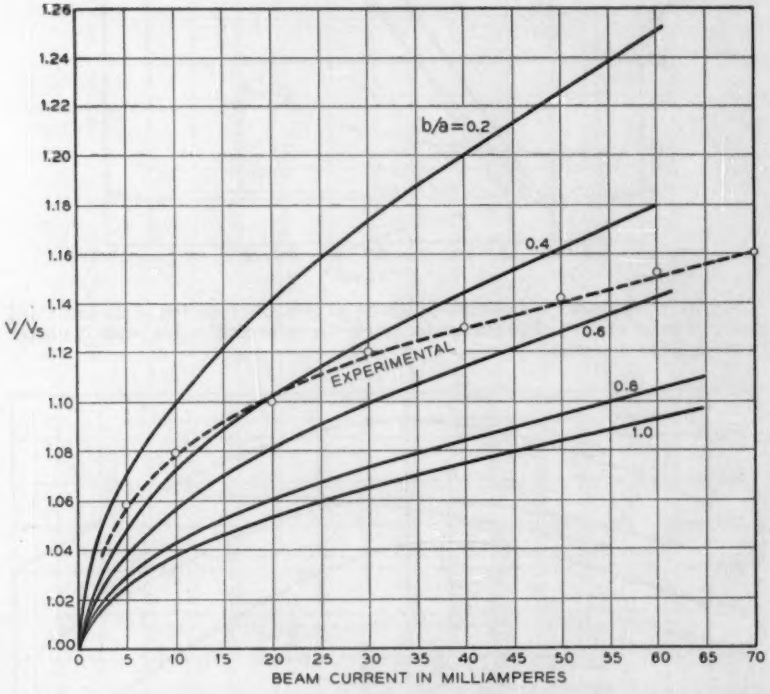


Fig. 34 — Measured and calculated ratio of voltage for maximum gain to synchronous voltage as a function of beam current. The calculated curves are shown for several values of the ratio of effective beam radius to mean helix radius (b/a). The location of the measured curve among the calculated ones is taken as an indication of the effective value of b/a in the M1789. At 40 ma it is about 0.5.

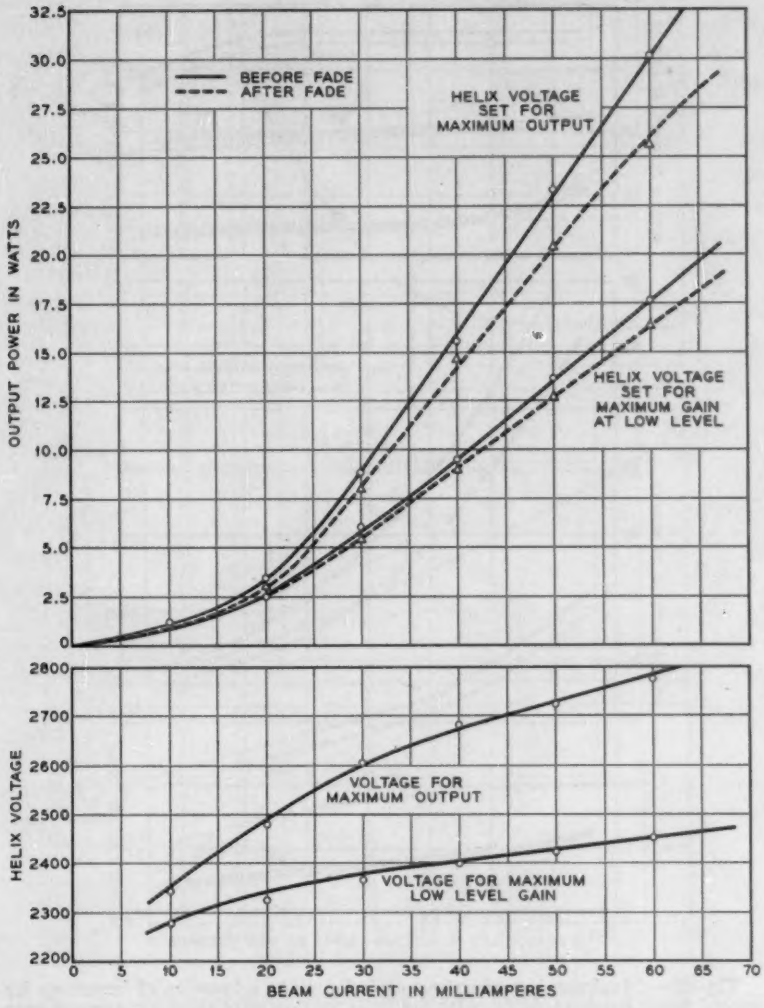


Fig. 35 — Maximum power output and helix voltage as functions of beam current. Curves are shown for before and after fading, and for the helix voltage adjusted for the maximum gain at low-level and for maximum output.

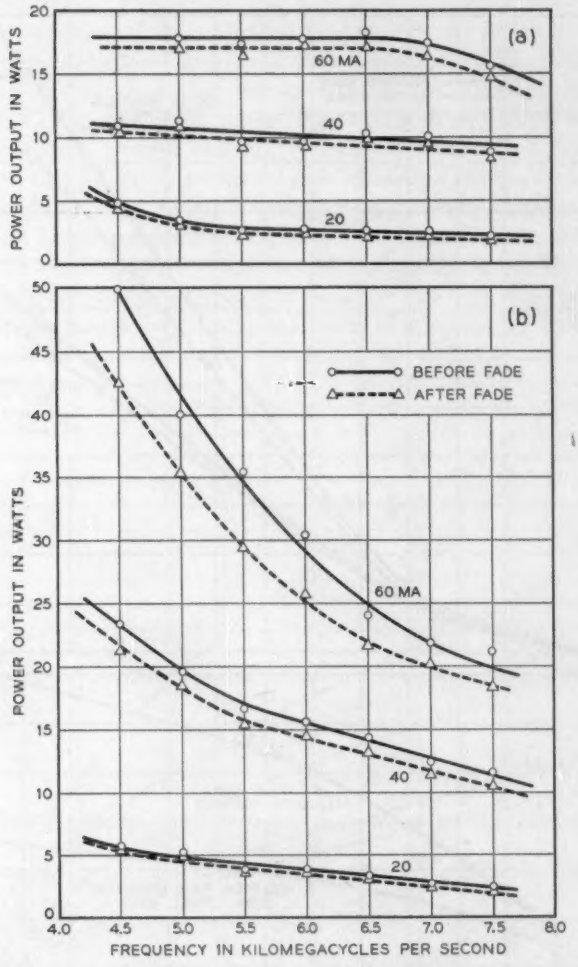


Fig. 36 — Maximum power output after fading as a function of frequency for several beam currents; in (a) with the helix voltage adjusted for maximum gain at low-level and in (b) with the helix voltage adjusted for maximum power output.

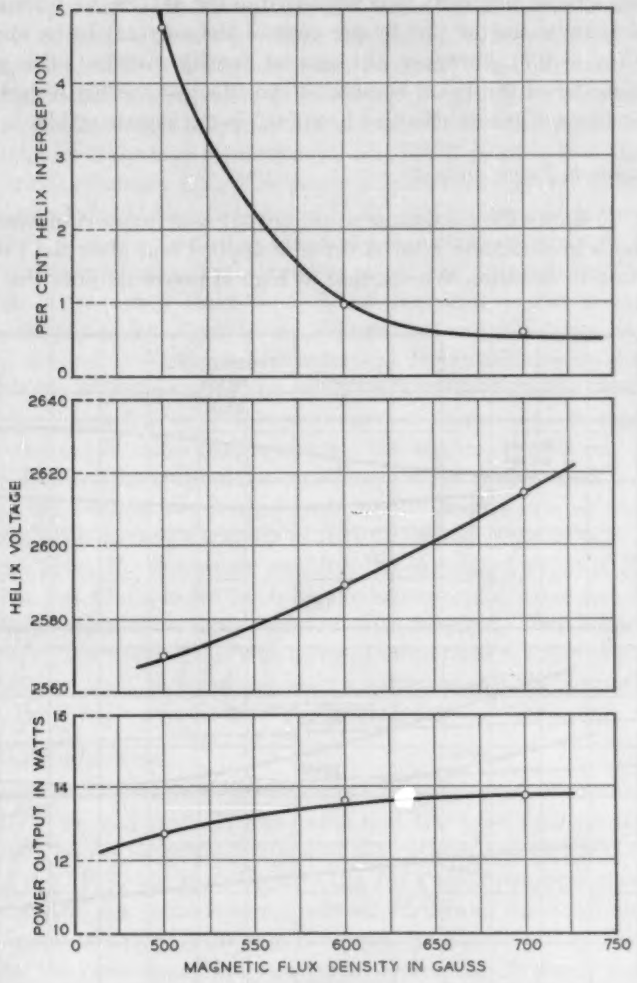


Fig. 37 — Maximum power output after fading, voltage for maximum output, and helix interception at maximum output as functions of magnetic flux density. These measurements were made using magnetic circuits charged to different strengths. Helix interception above about one per cent is undesirable if long tube life is required.

focusing system similar to that employed in the M1789, we estimate the actual beam diameter (for 99 per cent of the current) to be about 65 mils ($b/a = 0.7$). However, the current density distribution is peaked at the center of the beam because of the effect of thermal velocities of the electrons. Thus an effective b/a of 0.5 is not unreasonable.

Maximum Power Output

Fig. 35 shows the maximum power output as a function of beam current both immediately after rf drive is applied and after the tube has had time to stabilize. We see that at high rf power outputs the fading

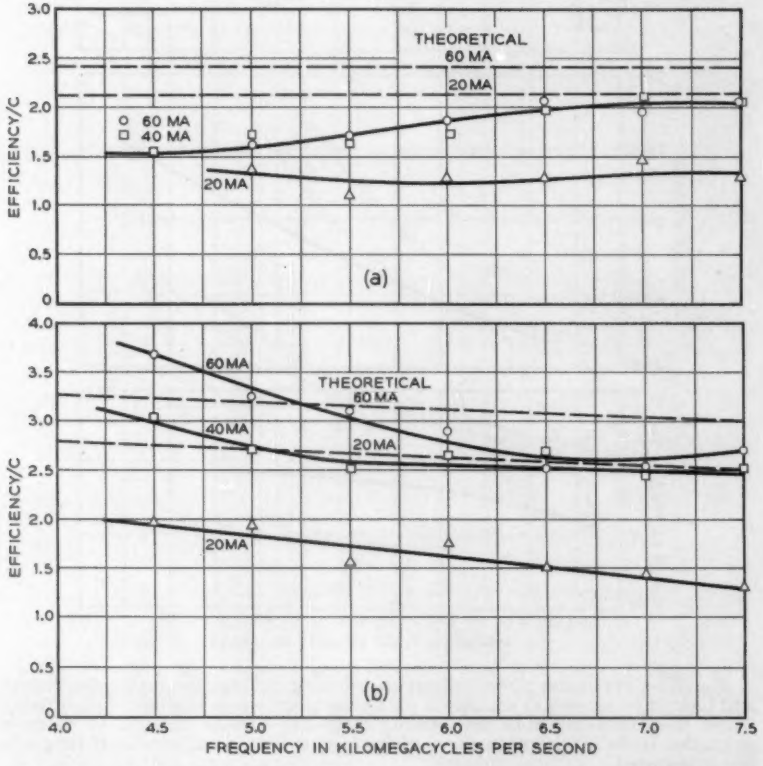


Fig. 38 — Ratio of electronic efficiency to gain parameter C as a function of frequency. The efficiencies used for this comparison are all before fading. The dotted lines are estimated from the Tien theory corrected for the intrinsic loss of the helix. The curves in (a) are for the case of the helix voltage adjusted for the maximum low-level gain and those in (b) for the case of the helix voltage adjusted for maximum power output.

becomes very serious and eventually limits the TWT output to about 30 watts. If it were necessary to reduce this fading, the envelope shrinking technique illustrated in Fig. 16 could be used. The maximum power output after fading is shown as a function of frequency for several beam currents in Fig. 36 and as a function of magnetic flux density in Fig. 37.

The theory of the high level behavior of a TWT[4] predicts that the ratio of electronic efficiency (i.e., E = power output/beam power) to the gain parameter C should be a function of C, QC and γb (where b is the beam diameter). However, with the range of parameters encountered in the M1789, the variation in E/C should be small. Fig. 38(a) shows E/C as a function of frequency when the TWT is operating at the voltage for maximum gain at low signal levels. Fig. 38(b) shows the maximum value of E/C obtainable at elevated helix voltage. In both figures we show the efficiency as estimated using the results of Tien[4] corrected for the effect of intrinsic loss following the procedure of Cutler and Brangaccio.[5] All efficiencies in these two figures are the electronic efficiency before fading. It would be quite difficult to compare the efficiency after fading with theory because the intrinsic attenuation in this case varies along the helix in an unknown manner so that we cannot properly take it into account. From the figures we see that the calculated value of E/C at 6,000 mc and 40 ma is not far from the experimental value but the experimental points show more variation with frequency than is predicted by theory. The low efficiency at 20 ma results from the fact that there is insufficient gain between the helix attenuator and the output. As a result, the TWT "overloads in the attenuation."

4.4 *Noise Performance*

A new and important noise phenomenon was observed in the course of the M1789 development. It was found that the noise figure is strongly dependent on the magnetic flux linking the cathode and on the rf output level of the TWT. For example, with the TWT operating near maximum output and with a cathode completely shielded from the magnetic field, noise figures of about 50 db were observed. By allowing 20 gauss at the cathode, the noise figure was reduced to 30 db. Fig. 39 shows the noise figure as a function of magnetic flux density at the cathode for several values of rf power output. We see that there is a peak of noise figure roughly symmetrical about zero flux at the cathode, and that the magnitude of this peak is considerably increased by operating the TWT at high output levels.

Some additional observed properties of the noise peak are:

(1) The magnitude depends on the synchronous voltage of the helix. For a 1,600-volt helix it is about 10 db higher than shown in Fig. 39 and

for a 2,600-volt helix it is about 5 db lower. The noise figure for 25 gauss at the cathode remains constant, however.

(2) There appears to be a threshold level of about 15-ma beam current below which the peak does not occur. Between 15 and 25 ma the peak increases. Above 25 ma it is roughly constant in magnitude.

(3) The peak can be considerably reduced by intercepting some of the edge electrons before they reach the helix region.

For this discussion it has been necessary to extend the concept of noise figure to the case of non-linear operation of the TWT. Essentially this noise figure is defined by the means we use to determine it. A block diagram of the equipment is shown in Fig. 40. The outputs of a calibrated broad band noise source and a signal oscillator are combined and used for the input to the TWT under test. The noise output from the TWT is passed through a filter tuned about 100 mc away from the signal so as to reject the carrier. It is then detected by a receiver tuned to the filter frequency. The noise figure is measured by turning the noise source off and on, noting the change in receiver output level and calculating the noise figure in the conventional manner. This procedure reduces to an ordinary noise figure measurement in the absence of input signal.

There are other ways that could be used to measure noise figure of a non-linear amplifier. A method more closely related to the use of the

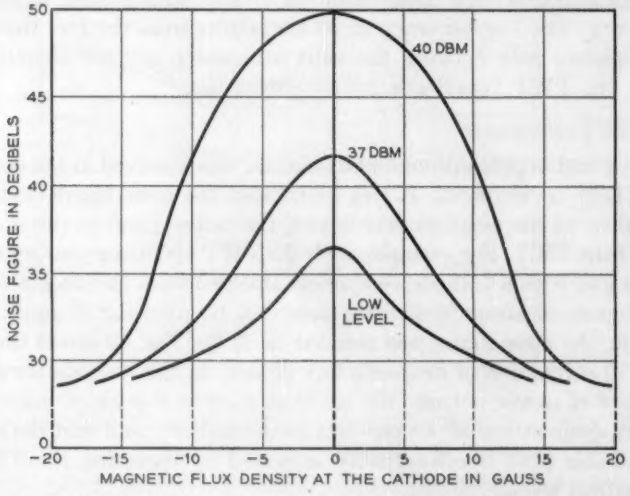


Fig. 39 — Noise figure as a function of magnetic flux density at the cathode for several values of rf power output. The flux density was varied by using an inductive heater through which ac current was passed. The present M1789 uses 19 gauss at the cathode, all of which is obtained from the focusing magnet — the heater now being non-inductive.

TWT in an FM radio relay was investigated briefly. In this measurement an FM receiver tuned to the carrier frequency was used to detect the noise modulation present in the TWT output. The noise figure was determined in the usual manner from the ratio of receiver outputs with the noise source turned off and on. When the TWT was operated in the linear region, this measurement gave the same result that our first method did. With the TWT operated in the non-linear region it gave a value within a few db of that obtained from the first method.

The cause of the high noise output observed for low magnetic flux densities at the cathode is at the present time not clearly understood. Fried at MIT and Ashkin and Rigrod at Bell Laboratories have all probed the beam formed by guns of the M1789 type and have found certain anomalous effects. Normally one would expect to find a standing wave of noise current along the electron beam. For the M1789 gun they find instead that after about two minima of the standing wave pattern, the noise current on the beam begins to grow and continues to do so until a saturation value is reached. The noise current at this saturation

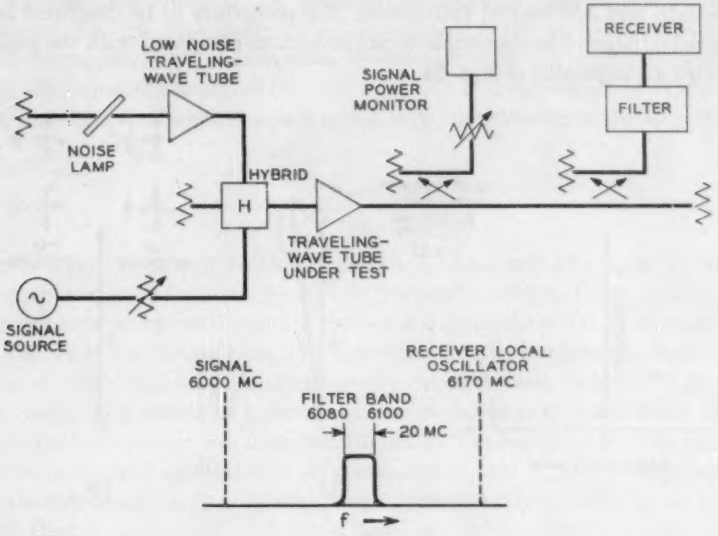


Fig. 40 — Block diagram of noise measuring equipment. The noise source consists of a fluorescent lamp the output of which is amplified by a low-noise TWT so as to bring the noise level to about 35 db above kTB at the M1789 input. The output from the M1789 is passed through a 20-mc bandpass filter which eliminates both the single frequency test signal and the noise in the image band of the receiver. The noise figure is measured by noting the difference in noise level at the receiver output with the noise source off and on, in a manner similar to that used in a conventional noise figure measurement.

value may be considerably higher than the original average noise level. As is the case with the noise figure in the M1789, the growing noise current has been found to be very sensitive to magnetic field at the cathode. By allowing sufficient field to link the cathode, the growing noise current can be eliminated leaving the normal noise current standing wave pattern on the beam. This phenomenon is not peculiar to the M1789 gun. It has been observed by various workers at MIT[6] and elsewhere on other guns producing beams with comparable current densities. A satisfactory explanation for it has not, at the time of this writing, been arrived at. It seems safe to say, however, that the growing noise current on the beam is the source of the high noise figures obtained in the M1789 when the cathode is completely shielded from the magnetic field.

4.5 *Intermodulation*

It has been found that certain intermodulation effects in the M1789 can be predicted from a knowledge of the compression and AM-to-PM conversion. Alternatively, these effects can be used to determine compression and AM-to-PM conversion. The procedure to be described has the advantage of being simple to implement as compared with the phase bridge arrangement of Fig. 24.

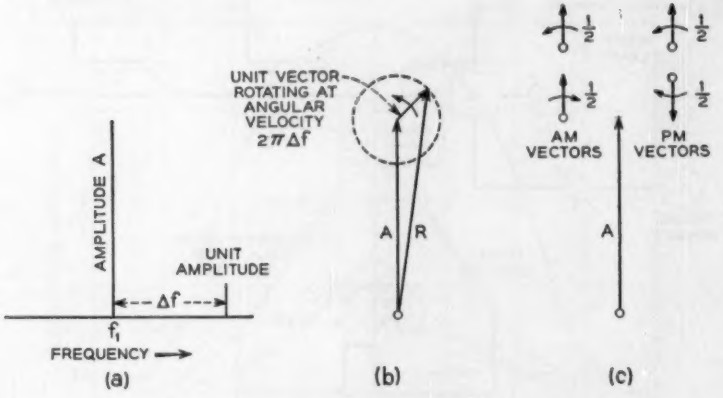


Fig. 41

(a) Spectrum of input signal to amplifier.

(b) Vector diagram of two input signals and the resultant signal (R) in a frame of reference rotating at an angular velocity $2\pi\Delta f$. Dotted line is the locus of the resultant signal.

(c) The rotating vector of the preceeding diagram can be broken down into a set of two vectors representing amplitude modulation and a set of two vectors representing frequency or phase modulation.

Intermodulation effects are ordinarily complicated and results are very hard to predict from single frequency measurements on an amplifier. For a TWT, however, one case — that in which two signals of very different amplitude are passed through the tube — can be treated simply. Consider an input to a TWT consisting of two signals at frequencies f_1 and $f_1 + \Delta f$ with the signal at f_1 being very much larger in amplitude. The composite signal applied to the amplifier will then be a signal at frequency f_1 which is amplitude and phase modulated at a rate Δf in an amount proportional to the relative magnitudes of the two signals. This can be represented vectorially as shown in Fig. 41(a) and b. In this figure the amplitude of the signal $f_1 + \Delta f$ has been normalized to unity. "A" thus represents the ratio of the larger to the smaller signal. The locus of the resultant signal is shown by the dotted line. The single rotating vector can be considered as the sum of vectors at $f_1 + \Delta f$ and $f_1 - \Delta f$ as shown in Fig. 41(c). One set of vectors produces PM and the other AM. The AM and PM vectors cancel at $f_1 - \Delta f$ and add at $f_1 + \Delta f$.

Suppose this signal is put through an amplifier operating in compression. For the time being let us assume this amplifier has no AM-to-PM conversion. The compression in the amplifier will operate on the AM sidebands of the signal but will leave the PM sidebands unaffected. Let us define the quantity c as a measure of compression in the amplifier by

$$c = 1 - \frac{\Delta V_0/V_0}{\Delta V_i/V_i} \tag{1}$$

where V_0 is the output voltage, V_i input voltage, and ΔV_0 is the change in output voltage for a change ΔV_i in the input voltage. This quantity is the per cent compression used in Section 4.2 divided by 100. If the signal in Fig. 41 is put through the amplifier while it is in compression, and the level of the signal at f_1 is subsequently brought back to amplitude A, we would then expect to have the situation shown in Fig. 42. Each AM sideband component has been multiplied by the factor (1-c). The locus of the composite signal is now elliptical. Let S_1 and S_2 be the magnitude of the sidebands at $f_1 + \Delta f$ and $f_1 - \Delta f$ respectively. From Fig. 42 it is seen that

$$S_1 = \tfrac{1}{2} + \tfrac{1}{2}(1 - c) = 1 - c/2 \tag{2}$$

$$S_2 = \tfrac{1}{2} - \tfrac{1}{2}(1 - c) = c/2 \tag{3}$$

When $c = 0$, the amplifier is operating in the linear region and $S_1 = 1$,

$S_2 = 0$. This is the condition in Fig. 41. When the amplifier is operating as a perfect limiter, $c = 1$ and $S_1 = S_2 = 0.5$. Thus, in this case, the sideband S_1 is down 6 db from its value when the amplifier is operating in the linear region.

When there is conversion of AM-to-PM in the amplifier, the situation becomes somewhat more complex. Suppose an AM signal is fed into the amplifier and that its voltage is given by

$$V = V_1(1 + \alpha \sin \omega_m t) \sin \omega_c t \tag{4}$$

where ω_c and ω_m are the carrier and modulating radian frequencies and V_1 and α are constants. The outputs will be given by

$$V = KV_1[1 + \alpha(1 - c) \sin \omega_m t] \sin (\omega_c t + k_p \alpha \sin \omega_m t) \tag{5}$$

Here K is the amplification, c is the compression factor and k_p is a factor which is a measure of the AM-to-PM conversion. It is seen that k_p is the output phase change for a given fractional input change α. Thus

$$k_p = \frac{\Delta\theta}{\alpha} \tag{6}$$

where $\Delta\theta$ is the phase change in radians caused by a fractional input change α. Later on it will be desired to express k_p in terms of degrees phase shift per db change in input amplitude. To express α in db we

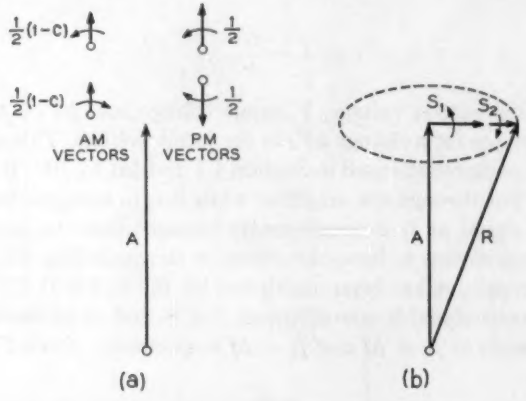


Fig. 42

(a) After passing through an amplifier in compression the AM sidebands are reduced in amplitude but the PM sidebands are unaffected. The lower two sidebands which represent a signal at frequency $f_1 - \Delta f$ no longer cancel and so there is a net signal at that frequency.

(b) The locus of the resultant signal now assumes an elliptical shape.

must evaluate 20 $\log_{10}(1+\alpha)$. The quantity $\log_e(1+\alpha)$ can be expanded in a series to give

$$\log_e(1+\alpha) = \alpha - \frac{1}{2}\alpha^2 + \frac{1}{3}\alpha^3 + \cdots .$$

As long as $\alpha \ll 1$, we can approximate it by taking only the first term of the above expression. Converting to the base ten and converting $\Delta\theta$ from radians as it appears in (6) to degrees, we find that

$$k_p = 0.152 \frac{\Delta\theta \text{ (in degrees)}}{\Delta \text{ input level (in db)}} \tag{7}$$

Now let us consider the case in which the signal of Fig. 41 is put through an amplifier having AM-to-PM conversion. Fig. 43 shows the vector picture of the resulting signal after the level of the signal at f_1 has been brought back to amplitude A. In this case the original PM sidebands and the compressed AM sidebands are the same as in Fig. 42, but there is now an additional set of PM sidebands as a result of the AM-to-PM conversion. Since the peak deviation of output phase due to this latter set of sidebands comes when the instantaneous amplitude is either a maximum or a minimum, they are 90 degrees out of phase with the other two sets of sidebands. From Fig. 43 it is seen that we can write

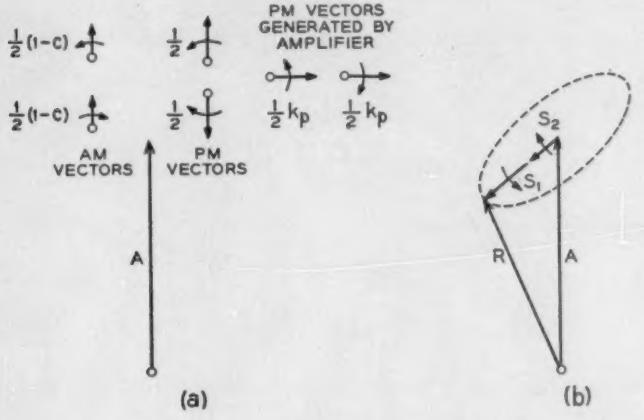


Fig. 43

(a) After passing through an amplifier having both compression and amplitude to phase conversion, the AM vectors are reduced in magnitude and a new set of PM vectors have appeared.

(b) The locus of the resultant signal of the vectors shown above is elliptical but the axis is tilted with respect to vector A.

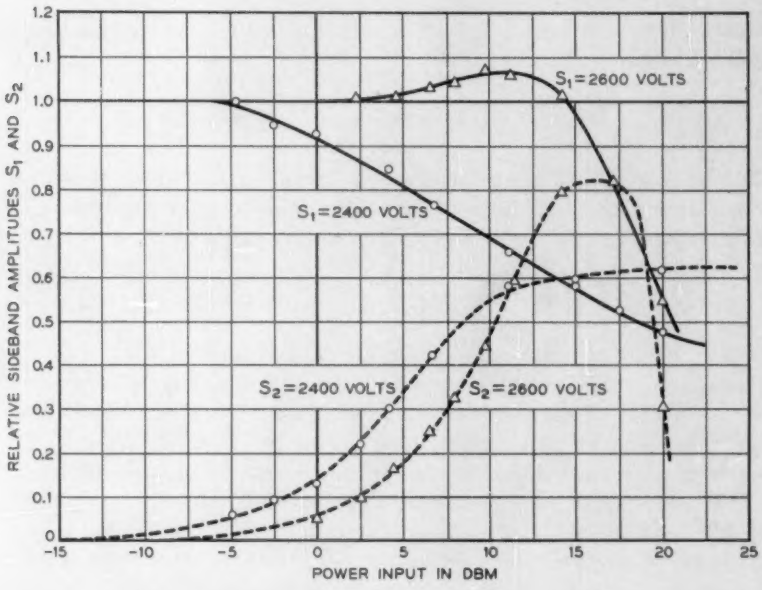


Fig. 44 — Relative side band amplitudes S_1 and S_2 for the M1789 as a function of power input for two values of helix voltage.

for the sideband amplitudes S_1 and S_2 at $f_1 + \Delta f$ and $f_1 - \Delta f$ respectively

$$S_1^2 = [\tfrac{1}{2} + \tfrac{1}{2}(1 - c)]^2 + \left[\frac{k_p}{2}\right]^2 = (1 - c/2)^2 + \left(\frac{k_p}{2}\right)^2 \quad (8)$$

$$S_2^2 = [\tfrac{1}{2} - \tfrac{1}{2}(1 - c)]^2 + \left[\frac{k_p}{2}\right]^2 = (c/2)^2 + \left(\frac{k_p}{2}\right)^2 \quad (9)$$

Solving for c and k_p we obtain

$$c = 1 - (S_1^2 - S_2^2) \quad (10)$$

$$k_p = 2\left[S_1^2 - \left(\frac{1 + S_1^2 - S_2^2}{2}\right)^2\right]^{1/2} \quad (11)$$

Thus we see that from a measurement of the amplitudes S_1 and S_2 the values of c and k_p can be determined.

To check the validity of this approach to intermodulation, we determined the values of compression and AM-to-PM conversion for an M1789 from an intermodulation measurement and compared them with values obtained using the phase bridge set-up described in Section 4.2. In the intermodulation measurement the two signals were 100 mc apart

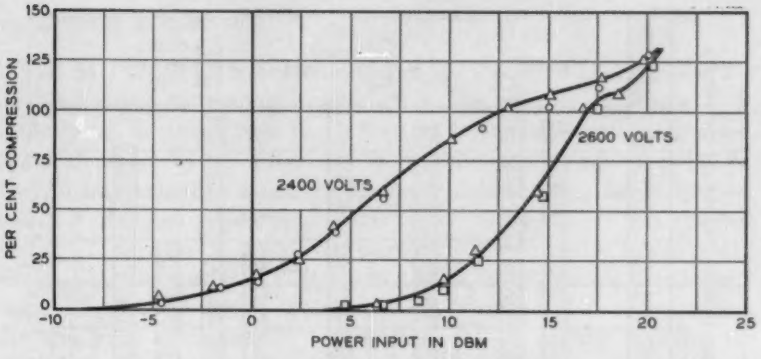


Fig. 45 — Compression as a function of input level for two values of helix voltage. Triangles represent data obtained with the test set of Fig. 24. Circles and squares represent data obtained by the two signal intermodulation measurement.

in frequency and 30 db different in level. From measurements of signal strength at the various frequencies involved, the magnitudes of S_1 and S_2 were determined with the results shown in Fig. 44. From these results the values of c and k_p were calculated and then converted to % compression and degrees per db in order to compare with the results of

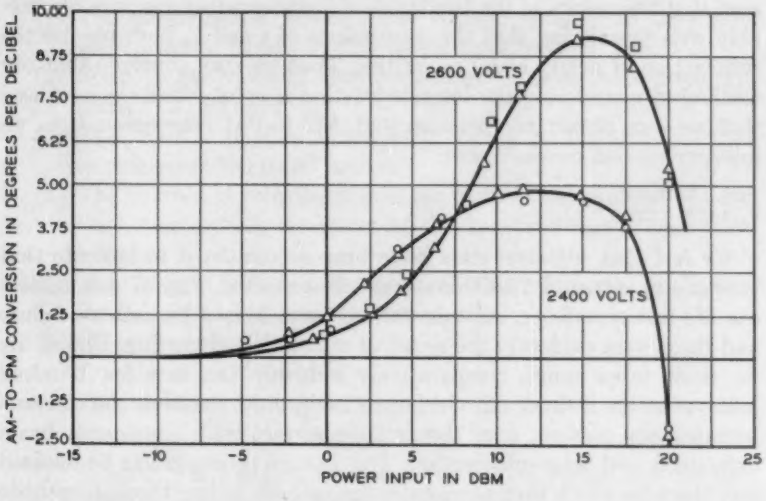


Fig. 46 — Conversion of amplitude modulation to phase modulation as a function of input level for two values of helix voltage. Triangles represent data obtained with the test set of Fig. 24. Circles and squares represent data obtained by the two signal intermodulation measurement.

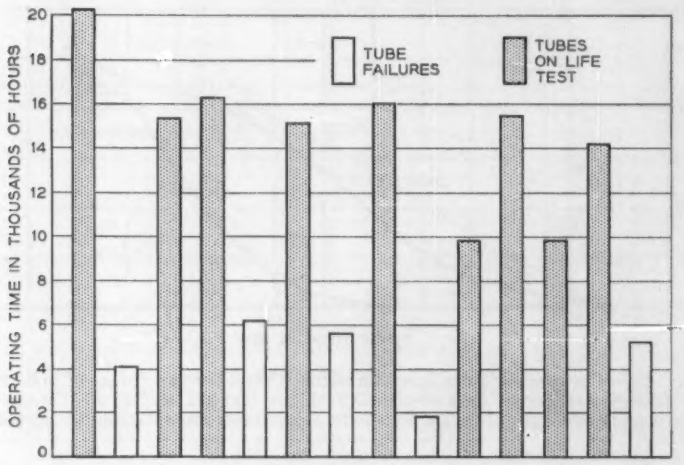


Fig. 47 — Life test results. The open bars indicate tubes that have failed; the solid bars tubes that were operating as of May 1, 1956. These tubes were operated with cathode temperatures between 720° and 760°C.

Figs. 23(c) and 23(e). The latter curves are repeated as Figs. 45 and 46 with the experimental points calculated from S_1 and S_2 shown. It is seen that the results of the two types of measurements compare remarkably well considering that the calculations of c and k_p both require the subtraction of nearly equal quantities. Thus we may conclude that our method of considering the intermodulation is substantially correct and that we can obtain compression and AM-to-PM conversion from an intermodulation measurement.

V. LIFE TESTS

We feel that sufficient data have been accumulated to indicate that tube life in excess of 10,000 hours can be expected. Fig. 47 summarizes our life test experience. All tube failures were caused by cathode failure and these were evidently the result of exhaustion of coating. End of life for these tubes comes comparatively suddenly i.e., in a few hundred hours after the cathode current begins to drop. At this time the emission becomes non-uniform over the cathode surface with consequent beam defocusing and helix interception. This in turn causes gas to be released into the tube which then accelerates the cathode failure through cathode poisoning. The rf performance remained good over the tube life — the gain and output power actually increasing slightly near the end of life as the beam started to defocus.

VI. ACKNOWLEDGMENTS

The M1789 TWT is the outcome of an intensive effort which has included many individuals in addition to the authors. R. Angle, J. S. Gellatly, E. G. Olson, and R. G. Voss all have contributed to the mechanical design of the tube and to its reduction to practice. R. W. DeVido has materially assisted with the electrical testing. M. G. Bodmer and J. F. Riley have been responsible for setting up the life test program and J. C. Irwin and J. A. Saloom contributed importantly to the design work on the electron gun. P. P. Cioffi and M. S. Glass have been largely responsible for the design of the magnetic circuits and P. I. Sandsmark for the helix-to-waveguide transducers. D. O. Melroy studied the effects of positive ions and performed the experiments on ion bombardment referred to in Section III. D. R. Jordan contributed to the studies on noise. In addition to the above, the authors would like to thank E. D. Reed for his very helpful criticism of this manuscript.

Appendix I — Gain Calculations

The gain calculations for the M1789 follow the procedure outlined by Pierce[7] with some minor modifications. The steps involved in the gain calculations for the loss free region of the helix are as follows:

(1) The experimental synchronous voltage is used to determine γa and the dielectric loading factor as defined by Tien.[8]

(2) From γa the value of helix impedance K is obtained from Appendix VI of Pierce.[7]

(3) The value of K is corrected using Tien's[8] results and C is then calculated in the usual manner.

(4) The number of wavelengths N_1 per inch of helix is obtained using the experimentally determined (from synchronous voltage) wavelength.

(5) The value of ω_q/ω is determined. In this calculation the curves for ω_p/ω_q from Watkins[9] are employed.

(6) QC is determined from

$$QC = \left(\frac{1}{2C}\,\frac{\omega_q}{\omega}\right)^2$$

(7) From QC, B is determined from Fig. 8.10 of Pierce[7] and the gain BCN_1 in the loss free region is calculated.

In calculating the effect of the attenuator section, we have had to make some rather gross assumptions. Fortunately, it turns out that the

gain in the attenuator is a small fraction of the total gain in the tube so that the over-all gain is not particularly sensitive to the means we use for treating the attenuator. Essentially what we have done is to consider the high loss part of the attenuator as a severed helix region and the low loss part of the attenuator as a lossy helix region.

Fig. 48 shows the value of the growing wave parameter as a function of the loss parameter d for various values of QC as calculated from theory. Because of discontinuity losses to the growing wave as it propagates in a region of gradually increasing loss, the actual gain will be less than that calculated from Fig. 48. Some rather crude probe measurements have indicated that the effective x vs. d curve can be approximated by a straight line through the $d = 0$ and $d = 1$ points — the dotted line in Fig. 48.

Since the helix is effectively severed by the high loss portion of the attenuator we must subtract some discontinuity loss from the gain in the attenuator region. The effective drift length in the severed region is unknown so this discontinuity loss cannot be accurately calculated from the low-level theory. The discussion in chapter nine of Pierce[7] indicates that an average value of about 6 db is reasonable.

An alternate method of treating the attenuator was also tried. In this calculation, the x vs. d curves in Fig. 48 were assumed to be correct to

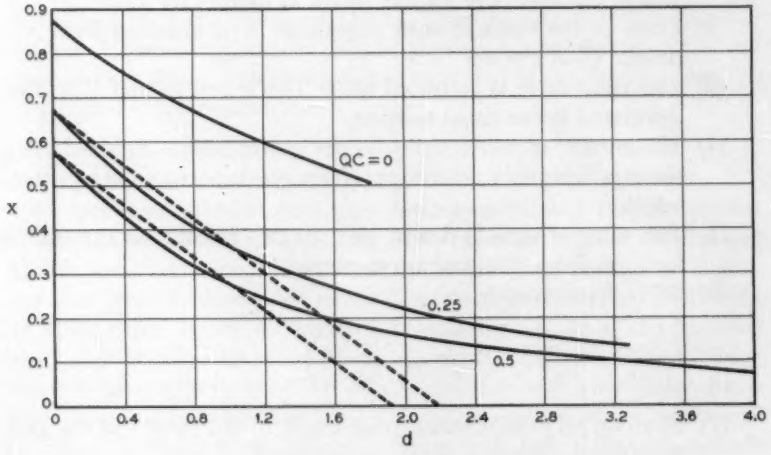


Fig. 48 — Curves of growing wave parameter x as a function of loss parameter d showing approximation (dotted lines) used in gain calculations for the M1789.

$d = 1$. The region for which $d > 1$ was considered as a severed helix region with 6-db discontinuity loss. Calculations using this procedure gave total gains for the TWT within a couple of db of the first method.

The remaining steps in calculating the gain of the TWT are therefore:

(8) The quantity α is determined from the slope of the dotted lines in Fig. 48.

(9) The length of helix, ℓ_e in the attenuator for which $x > 0$ is determined by using Fig. 48.

(10) The total attenuation L, in the section of the attenuator effective in producing gain is calculated.

(11) The initial loss parameter A is obtained from Fig. 94 of Pierce.[7]

(12) The gain is calculated from

$$\text{Gain} = A - 6\text{db} + \alpha L + BCN_1 (3.5 + \ell_e)$$

where the six db is the discontinuity loss in the attenuator section and the 3.5 inches is the length of loss free helix.

Glossary of Symbols

α loss factor from Pierce[7]
A discontinuity loss parameter at input of helix from Pierce[7]
B magnetic flux density or the space charge parameter from Pierce[7]
B_B Brillouin flux density for a beam entirely filling the helix
C gain parameter from Pierce[7]
a helix radius
b beam radius
d loss parameter from Pierce[7]
f frequency
I_k cathode current
I_a accelerator current
I_h helix current
I_c collector current
k $2\pi/\lambda_0$ where λ_0 is the free space wavelength
ℓ_e length of helix attenuator in which gain is possible
L loss in the part of the attenuator section which is capable of producing gain.
N number of wavelengths in TWT
N_1 number of wavelengths on the helix per inch
QC space charge parameter from Pierce[7]
$\overline{r_a}$ anode radius of curvature of gun
$\overline{r_c}$ cathode radius of curvature of gun
r_{min} minimum beam radius from Pierce[10]

r_c cathode radius
r_{95} radius at the beam minimum through which 95 per cent of the current flows
σ standard deviation of electron trajectory
T_k cathode temperature
V_a accelerator voltage
V_h helix voltage
V_c collector voltage
x growing wave parameter from Pierce
ω radian frequency
ω_c carrier radian frequency
ω_m modulating signal radian frequency
ω_p radian plasma frequency
ω_q corrected radian plasma frequency
c compression factor
k_p AM-to-PM conversion factor
γ radial propagation constant

References

1. Cutler, C. C., Spurious Modulation of Electron Beams, Proc. I.R.E., **44**, pp. 61–64, Jan., 1956.
2. Danielson, W. E., Rosenfeld, J. L., and Saloom, J. A., A Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type, B.S.T.J. **35**, pp. 375–420, March, 1956.
3. Augustine, C. F., and Slocum, A., 6KMC Phase Measurement System For Traveling-Wave Tubes, I.R.E. Trans. **PGI-4**, Oct., 1955.
4. Tien, P. K., A Large Signal Theory of Traveling-Wave Amplifiers, B.S.T.J., **35**, pp. 349–374, March, 1956.
5. Brangaccio, D. J., and Cutler, C. C., Factors Affecting Traveling-Wave Tube Power Capacity, I.R.E. Trans. **PGED-3**, June, 1953.
6. Smullin, L. D., and Fried, C., Microwave Noise Measurements on Electron Beams, I.R.E. Trans., **PGED-4**, Dec., 1954.
7. Pierce, J. R., Traveling-Wave Tubes, D. Van Nostrand, Inc., 1950.
8. Tien, P. K., Traveling-Wave Tube Helix Impedance, Proc. I.R.E., **41**, pp. 1617–1623, Nov., 1953.
9. Watkins, D. A., Traveling-Wave Tube Noise Figure, Proc. I.R.E., **40**, pp. 65–70, Jan., 1952.
10. Pierce, J. R., Theory and Design of Electron Beams, D. Van Nostrand, Inc., 1949.

Helix Waveguide

By S. P. MORGAN and J. A. YOUNG

(Manuscript received July 23, 1956)

Helix waveguide, composed of closely wound turns of insulated copper wire covered with a lossy jacket, shows great promise for use as a communication medium. The properties of this type of waveguide have been investigated using the sheath helix model. Modes whose wall currents follow the highly conducting helix have attenuation constants which are essentially the same as for copper pipe. The other modes have very large attenuation constants which depend upon the helix pitch angle and the electrical properties of the jacket. Approximate formulas are given for the propagation constants of the lossy modes. The circular electric mode important for long-distance communication has low loss for zero-pitch helices. The propagation constants of some of the lossy modes in helix waveguide of zero pitch have been calculated numerically, as functions of the jacket parameters and the guide size, in regions where the approximate formulas are no longer valid. Under certain conditions the attenuation constant of a particular mode may pass through a maximum as the jacket conductivity is varied.

Glossary of symbols

a	Inner radius of waveguide
$h = \beta - i\alpha$	Complex phase constant
n	Angular mode index
p	Denotes p_{nm} or p_{nm}' according to context
p_{nm}	m^{th} zero of $J_n(x)$
p_{nm}'	m^{th} zero of $J_n'(x)$
r, θ, z	Right-handed cylindrical coordinates
α	Attenuation constant
β	Phase constant
$\beta_0 = 2\pi/\lambda_0 = \omega(\mu_0\epsilon_0)^{1/2}$	Free-space phase constant
ϵ_0	Permittivity of interior medium
ϵ	Permittivity of exterior medium
ϵ'	ϵ/ϵ_0
ϵ''	$\sigma/\omega\epsilon_0$

ζ_1	$[\omega^2\mu_0\epsilon_0 - h^2]^{1/2}$
ζ_2	$[\omega^2\mu_0\epsilon_0(\epsilon' - i\epsilon'') - h^2]^{1/2}$
λ_0	Free-space wavelength
$\lambda_c = 2\pi a/p$	Cutoff wavelength
μ_0	Permeability of interior and exterior media
$\nu = \lambda_0/\lambda_c = p\lambda_0/2\pi a$	Cutoff ratio
$\xi + i\eta$	$\dfrac{(\epsilon' - 1 + \nu^2 - i\epsilon'')^{1/2}}{\epsilon' - i\epsilon''}$
Π	Electric Hertz vector
Π^*	Magnetic Hertz vector
σ	Conductivity of exterior medium
ψ	Pitch angle of helix
ω	Angular frequency
$e^{i\omega t}$	Harmonic time dependence assumed throughout
$J_n(x)$	Bessel function of the first kind
$J_n'(x)$	$dJ_n(x)/dx$
$H_n^{(2)}(x)$	Hankel function of the second kind
$H_n^{(2)'}(x)$	$dH_n^{(2)}(x)/dx$

MKS rationalized units are employed throughout. Superscripts i and e are used to indicate the interior and exterior regions.

I. INTRODUCTION AND SUMMARY

Propagation of the lowest circular electric mode (TE_{01}) in cylindrical pipe waveguide holds great promise for low-loss long distance communication.[1, 2] For example, the TE_{01} mode has a theoretical heat loss of 2 db/mile in waveguide of diameter 6 inches at a frequency of 5.5 kmc/s, and the loss decreases with increasing frequency. Increased transmission bandwidth, reduced delay distortion, and reduced waveguide size for a given attenuation are factors favoring use of the highest practical frequency of operation. An increased number of freely propagating modes and smaller mechanical tolerances are the associated penalties. Any deviation of the waveguide from a straight circular cylinder gives rise to signal distortions because of mode conversion-reconversion effects.

One solution to mode conversion-reconversion problems is to obtain a waveguide having the desired low attenuation properties of the TE_{01} mode in metallic cylindrical waveguide and very large attenuation for all other modes, the unwanted modes.[1, 3] The low loss of the circular electric modes in ordinary round guide is the result of having only cir-

[1] S. E. Miller, B.S.T.J., **33**, pp. 1209–1265, 1954.
[2] S. E. Miller and A. C. Beck, Proc. I.R.E., **41**, pp. 348–358, 1953.
[3] S. E. Miller, Proc. I.R.E., **40**, pp. 1104–1113, 1952.

cumferential current flow at the boundary wall. All other modes in round guide have a longitudinal current present at the wall. Thus the desired attenuation properties can be obtained by providing a highly conducting circumferential path and a resistive longitudinal path for the wall currents. This is done in the spaced-disk line by sandwiching lossy layers between coaxially arranged annular copper disks.[4] Another possibility which has been suggested is a helix having a small pitch.

Helix waveguide, formed by winding insulated wire on a removable mandrel and coating the helix with lossy material, has been made at the Holmdel Radio Research Laboratory. Wires of various cross sections and sizes have been used to wind helices varying from $\frac{7}{16}$ to 5 inches in diameter, which have been tested at frequencies from 9 to 60 kmc/s. Pitch angles of from nearly 0° (wire in a plane perpendicular to the axis of propagation) to 90° (wire parallel to the axis of propagation) have been used. The helices having the highest attenuation for the unwanted modes while maintaining low loss for the TE_{01} mode are those wound with the smallest pitch from insulated wire of diameter 10 to 3 mils (American Wire Gauge Nos. 30 to 40). The high attenuation properties for unwanted modes also depend markedly on the electrical properties of the jacket surrounding the helix.

In this paper the normal modes of helix waveguide are determined using the sheath helix approximation, a mathematical model in which the helical winding is replaced by an anisotropic conducting sheath. A brief formulation of the boundary value problem leads to an equation which determines the propagation constants of modes in the helix guide. Since the equation is not easy to solve numerically, approximations are presented which show the effects of the pitch angle, the diameter, the conductivity and dielectric constant of the jacket, and the wavelength, when the conductivity of the jacket is sufficiently high.

By proper choice of the pitch angle and, in some instances, of the polarization, a helix waveguide can be made to propagate any mode of ordinary round guide, with an attenuation constant which should be essentially the same as in solid copper pipe. The pitch is chosen so that the wall currents associated with the desired mode follow the direction of the conducting wires. The losses to the other modes are in general much higher, and are determined by both the pitch angle and the jacket material.

Special attention is given in the present work to the limiting case of a helix of zero pitch, since the attenuation constant of the TE_{01} mode will be smallest when the pitch angle is as small as possible. To explore the

[4] Reference 3, p. 1111.

region where the approximate formulas for the propagation constants of the lossy modes break down, some numerical results have been obtained for helices of zero pitch using an IBM 650 magnetic drum calculator. Tables and curves are given showing the propagation constants of various modes in such a waveguide, as functions of the electrical properties of the jacket and for three different ratios of radius/wavelength. In many cases it is found that the attenuation constant of a given mode passes through a maximum as the jacket conductivity is varied, the other parameters remaining fixed. The numerical calculations indicate that it is possible to get unwanted mode attenuations several hundred to several hundred thousand times greater than the TE_{01} attenuation for the size waveguide that looks most promising for low-loss communication.

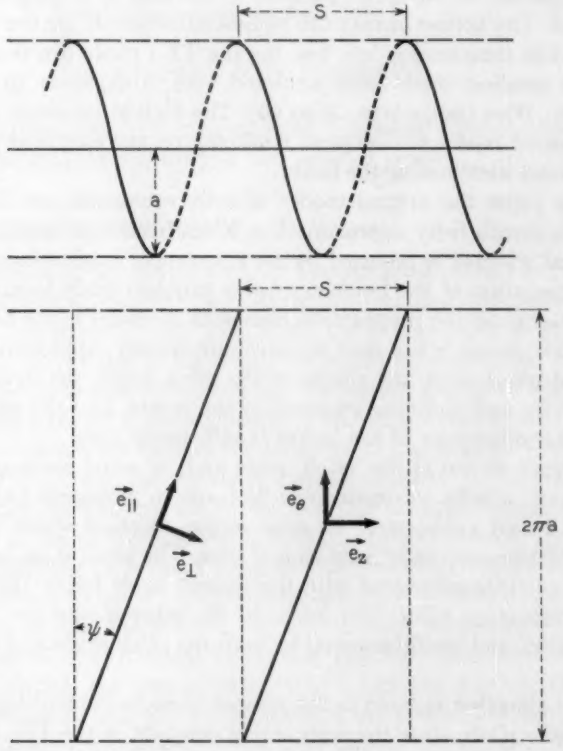


Fig. 1 — Schematic diagrams of the helical sheath and the helical sheath developed, showing the unit vectors and the periodicity.

II. SHEATH HELIX BOUNDARY VALUE PROBLEM

Ordinary cylindrical waveguide consists of a circular cylinder of radius a, infinite length, and zero (or very small) conductivity, imbedded in an infinite* homogeneous conducting medium. The sheath helix waveguide has the same configuration plus the additional property that at radius a dividing the two media, there is an anisotropic conducting sheath which conducts perfectly in the helical direction and does not conduct in the perpendicular direction. The attenuation and phase constants are determined by solving Maxwell's equations in cylindrical coordinates and matching the electric and magnetic fields at the wall of the guide.

The helix of radius a and pitch angle $\psi = \tan^{-1} s/2\pi a$ is shown in the upper part of Fig. 1. The developed helix as viewed from the inside when cut by a plane of constant θ and unrolled is shown in the lower part of the illustration. A new set of unit vectors $\vec{e}_\parallel$ and $\vec{e}_\perp$ parallel and perpendicular respectively to the helix direction is introduced. These are related to $\vec{e}_r$, $\vec{e}_\theta$, and $\vec{e}_z$ by

$$\vec{e}_r \times \vec{e}_\parallel = \vec{e}_\perp$$
$$\vec{e}_\parallel = \vec{e}_z \sin\psi + \vec{e}_\theta \cos\psi$$
$$\vec{e}_\perp = \vec{e}_z \cos\psi - \vec{e}_\theta \sin\psi$$

The boundary conditions at $r = a$ are

$$E_\parallel^{\,i} = E_\parallel^{\,e} = 0$$
$$E_\perp^{\,i} = E_\perp^{\,e}$$
$$H_\parallel^{\,i} = H_\parallel^{\,e}$$

where the superscript i refers to the interior region, $0 \leqq r \leqq a$, and the superscript e refers to the exterior region, $a \leqq r \leqq \infty$. An equivalent set of boundary conditions in terms of the original unit vectors is

$$\begin{aligned} E_z^{\,i} \tan\psi + E_\theta^{\,i} &= 0 \\ E_z^{\,e} \tan\psi + E_\theta^{\,e} &= 0 \\ E_z^{\,i} &= E_z^{\,e} \\ H_z^{\,i} \tan\psi + H_\theta^{\,i} &= H_z^{\,e} \tan\psi + H_\theta^{\,e} \end{aligned} \tag{1}$$

We are looking for solutions which are similar to the modes of or-

* The assumption of an infinite external medium is made to simplify the mathematics. The results will be the same as for a finite conducting jacket which is thick enough so that the fields at its outer surface are negligible.

dinary waveguide, i.e., "fast" modes as contrasted with the well-known "slow" modes used in traveling-wave tubes.[5, 6] To solve the problem we follow the procedure set up by Stratton[7] for the ordinary cylindrical waveguide boundary problem. The fields $\vec{E}$ and $\vec{H}$ are derived from an electric Hertz vector $\vec{\Pi}$ and a magnetic Hertz vector $\vec{\Pi}^*$ by

$$\begin{aligned} \vec{E} &= \vec{\nabla} \times \vec{\nabla} \times \vec{\Pi} - i\omega\mu\vec{\nabla} \times \vec{\Pi}^* \\ \vec{H} &= (\sigma + i\omega\epsilon)\vec{\nabla} \times \vec{\Pi} + \vec{\nabla} \times \vec{\nabla} \times \vec{\Pi}^* \end{aligned} \tag{2}$$

where

$$\begin{aligned} \vec{\Pi} &= \vec{e}_z \Pi_z \\ \vec{\Pi}^* &= \vec{e}_z \Pi^* \end{aligned} \tag{3}$$

and, assuming a time dependence exp $(i\omega t)$,

$$\begin{aligned} \Pi_z^i &= \sum_{n=-\infty}^{\infty} a_n{}^i J_n(\zeta_1 r) e^{-ihz-in\theta} \\ \Pi_z^e &= \sum_{n=-\infty}^{\infty} a_n{}^e H_n^{(2)}(\zeta_2 r) e^{-ihz-in\theta} \\ \Pi_z^{*i} &= \sum_{n=-\infty}^{\infty} b_n{}^i J_n(\zeta_1 r) e^{-ihz-in\theta} \\ \Pi_z^{*e} &= \sum_{n=-\infty}^{\infty} b_n{}^e H_n^{(2)}(\zeta_2 r) e^{-ihz-in\theta} \end{aligned} \tag{4}$$

In these expressions

$$\begin{aligned} \zeta_1^2 &= \omega^2\mu_0\epsilon_0 - h^2 \\ \zeta_2^2 &= \omega^2\mu_0\epsilon_0(\epsilon' - i\epsilon'') - h^2 \\ \epsilon' - i\epsilon'' &= \epsilon/\epsilon_0 - i\sigma/\omega\epsilon_0 \end{aligned}$$

where the interior region is assumed to have permittivity ϵ_0 and permeability μ_0, while the exterior region has permittivity ϵ, permeability μ_0, and conductivity σ. The superscripts i and e refer to the interior and exterior regions respectively, and the a's and b's are amplitude coefficients.

[5] J. R. Pierce, Proc. I.R.E., **35**, pp. 111–123, 1947.

[6] S. Sensiper, Electromagnetic Wave Propagation on Helical Conductors, Sc.D. thesis, M.I.T., 1951. In Appendix B of this reference, Sensiper shows that when the interior and exterior media are the same, only slow waves will exist except in special cases. Fast guided waves become possible if the conductivity of the exterior medium is sufficiently high.

[7] J. A. Stratton, Electromagnetic Theory, McGraw-Hill, New York, 1941, pp. 524–527. Note that Stratton uses the time dependence exp $(-i\omega t)$.

Attention is restricted to waves traveling in the positive z-direction, which are represented by the factor $\exp(-ihz)$, where $h\ (=\beta - i\alpha)$ is the complex phase constant. However it is necessary to consider both right and left circularly polarized waves; this accounts for the use of both positive and negative values of n.

Substitution of (2), (3), and (4) into the boundary conditions (1) leads to the following set of equations:

$$\left[\zeta_1^2 \tan\psi - \frac{hn}{a}\right] J_n(\zeta_1 a) a_n^{\,i} + i\omega\mu_0\zeta_1 J_n'(\zeta_1 a) b_n^{\,i} = 0$$

$$\left[\zeta_2^2 \tan\psi - \frac{hn}{a}\right] H_n^{(2)}(\zeta_2 a) a_n^{\,e} + i\omega\mu_0\zeta_2 H_n^{(2)\prime}(\zeta_2 a) b_n^{\,e} = 0$$

$$\zeta_1^2 J_n(\zeta_1 a) a_n^{\,i} - \zeta_2^2 H_n^{(2)}(\zeta_2 a) a_n^{\,e} = 0 \tag{5}$$

$$-i\omega\epsilon_0\zeta_1 J_n'(\zeta_1 a) a_n^{\,i} + \left[\zeta_1^2 \tan\psi - \frac{hn}{a}\right] J_n(\zeta_1 a) b_n^{\,i}$$

$$+ (\sigma + i\omega\epsilon)\zeta_2 H_n^{(2)\prime}(\zeta_2 a) a_n^{\,e} - \left[\zeta_2^2 \tan\psi - \frac{hn}{a}\right] H_n^{(2)}(\zeta_2 a) b_n^{\,e} = 0$$

If the conductivity of the exterior region is infinite, it is possible to satisfy the boundary conditions with only one of the amplitude coefficients different from zero; for example

$$\begin{aligned} b_n^{\,i} = a_n^{\,e} = b_n^{\,e} &= 0 \qquad & a_n^{\,i} = a_n^{\,e} = b_n^{\,e} &= 0 \\ a_n^{\,i} &\neq 0 \quad \text{or} & b_n^{\,i} &\neq 0 \\ J_n(\zeta_1 a) &= 0 & J_n'(\zeta_1 a) &= 0 \end{aligned}$$

The first case corresponds to TM modes and the second to TE modes in a perfectly conducting circular guide. Linearly polarized modes may be represented as combinations of terms in $a_n^{\,i}$ and $a_{-n}^{\,i}$, or $b_n^{\,i}$ and $b_{-n}^{\,i}$.

If the exterior region is not perfectly conducting, one can still find solutions having the fields confined to the interior region by properly choosing the angle of the perfectly conducting helical sheath. For example, it is easy to verify that equations (5) are satisfied under the following conditions:

$$\begin{aligned} a_n^{\,i} = a_n^{\,e} = b_n^{\,e} &= 0 \\ b_n^{\,i} &\neq 0 \\ \tan\psi &= \frac{hn}{\zeta_1^2 a} \\ J_n'(\zeta_1 a) &= 0 \end{aligned}$$

If $n \neq 0$, these conditions correspond to circularly polarized TE_{nm} waves, in which the wall currents follow the direction of the conducting sheath. If $n = 0$, then $\psi = 0$, and one has TE_{0m} modes with circumferential currents only.

The equations can also be satisfied with

$$b_n^{\,i} = a_n^{\,e} = b_n^{\,e} = 0$$
$$a_n^{\,i} \neq 0$$
$$\psi = 90°$$
$$J_n(\zeta_1 a) = 0$$

corresponding to the TM_{nm} modes (either circularly or linearly polarized) of a perfectly conducting pipe, which are associated with longitudinal wall currents only.

In the general case when the jacket is not perfectly conducting and the helix pitch angle is not restricted to special values, it is necessary to solve (5) simultaneously for the field amplitudes. The equations admit a nontrivial solution if and only if the determinant of the coefficients of the a's and b's vanishes. The transcendental equation which results from equating the determinant of the coefficients to zero is

$$\zeta_2\left[\left(\zeta_1 \tan\psi - \frac{hn}{\zeta_1 a}\right)^2 \frac{J_n(\zeta_1 a)}{J_n'(\zeta_1 a)} - \omega^2\mu_0\epsilon_0 \frac{J_n'(\zeta_1 a)}{J_n(\zeta_1 a)}\right] \tag{6}$$
$$= \zeta_1\left[\left(\zeta_2 \tan\psi - \frac{hn}{\zeta_2 a}\right)^2 \frac{H_n^{(2)}(\zeta_2 a)}{H_n^{(2)\prime}(\zeta_2 a)} - \omega^2\mu_0\epsilon_0(\epsilon' - i\epsilon'') \frac{H_n^{(2)\prime}(\zeta_2 a)}{H_n^{(2)}(\zeta_2 a)}\right]$$

The solution of this equation determines the propagation constant ih and therefore the attenuation and phase constants α and β. When ih has been obtained, it is a straightforward matter to determine the a and b coefficients from equations (5) and the electric and magnetic fields from (2), (3), and (4).

It is well known[8] that the only pure TE or TM modes that can exist in a circular waveguide with walls of finite conductivity are the circularly symmetric TE_{0m} and TM_{0m} modes. The other modes are all mixed modes whose fields are not transverse with respect to either the electric or the magnetic vector. In general the modes of helix waveguide are also mixed modes, and no entirely satisfactory scheme for labeling them has been proposed. In the present paper we shall call the modes TE_{nm} or TM_{nm} according to the limits which they approach as the jacket conductivity becomes infinite, even though they are no longer transverse and their

[8] Reference 7, p. 526.

field patterns may be quite different when the jacket is lossy. This system is not completely unambiguous, because as will appear in Section IV the mode designations thus obtained are not always unique. However it is a satisfactory way to identify the modes so long as the jacket conductivity is high enough for the loss to be treated as a perturbation. Approximations derived on this basis are presented in the next section.

III. APPROXIMATE EXPRESSIONS FOR PROPAGATION CONSTANTS

If the jacket were perfectly conducting, the helix waveguide modes would be the same as in an ideal circular waveguide, with propagation constants given by

$$ih = i\beta_{nm} = i(2\pi/\lambda_0)(1 - \nu^2)^{1/2}$$

where

$\nu = \lambda_0/\lambda_e = p\lambda_0/2\pi a$

$p = m^{\text{th}}$ zero of $J_n(x)$ for TM_{nm} mode, or m^{th} zero of $J_n'(x)$ for TE_{nm} mode

If the jacket conductivity is sufficiently large, approximate solutions of (6) may be found by replacing $H_n^{(2)}(\zeta_2 a)$ and $H_n^{(2)\prime}(\zeta_2 a)$ with their asymptotic expressions, and expanding $J_n(\zeta_1 a)$ or $J_n'(\zeta_1 a)$ in a Taylor series near a particular zero. This calculation is carried out in the appendix. The propagation constant may be written in the form

$$ih = \alpha + i(\beta_{nm} + \Delta\beta)$$

where to first order the perturbation terms are

TM_{nm} *modes*

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}} \frac{1}{1 + \tan^2 \psi} \tag{7a}$$

TE_{nm} *modes*

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}} \frac{\nu^2 p^2}{p^2 - n^2} \frac{[\tan \psi - n(1 - \nu^2)^{1/2}/p\nu]^2}{1 + \tan^2 \psi} \tag{7b}$$

and

$$\xi + i\eta = (\epsilon' - i\epsilon'')^{-1/2}$$

$$\epsilon' = \epsilon/\epsilon_0 , \qquad \epsilon'' = \sigma/\omega\epsilon_0$$

The approximations made in deriving (7) are discussed in the appen-

dix. In practice, the range of validity of these expressions is usually limited by the criterion

$$\frac{a(1-\nu^2)^{1/2}}{\nu}\,|\,\alpha + i\Delta\beta\,| \ll 1 \tag{8}$$

The numerical calculations described in Section IV indicate that the approximations are good so long as the left-hand side of (8) is less than about 0.1, and that they break down a little sooner for TE modes than for TM modes.

Inspection of (7) reveals three cases of particular interest, namely $\psi = 0°$, $\psi = \tan^{-1} n(1-\nu^2)^{1/2}/p\nu$, and $\psi = 90°$. These cases, which were mentioned in Section II and are discussed again below, correspond to preferential propagation of certain modes, in which the wall currents follow the direction of the conducting helix. The preferred modes have zero attenuation in the present treatment because the helical sheath is assumed to be perfectly conducting. In practical helices wound from insulated copper wire the loss should be only slightly greater than in round copper pipe of the same diameter. The slight increase (of magnitude 10 per cent to 30 per cent) is due to the slightly nonuniform current distribution in the wires, an effect that can be kept small by keeping the gaps between the wires of the helix small. In general the attenuation constants of modes whose wall currents do not follow the helix are orders of magnitude larger than the attenuation constants of the preferred modes.

$\psi = 0°$

The circular electric (TE_{0m}) modes have attenuation constants substantially the same as in solid copper pipe. The additional TE_{0m} loss if the pitch angle is not quite zero is proportional to $\tan^4 \psi$. This added loss can be made very small by using fine wire for winding the helix.

The losses for the unwanted modes can be made large by a proper choice of jacket material. When $\psi = 0$, equations (7) yield

TM_{nm} *modes*

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1-\nu^2)^{1/2}} \tag{9a}$$

TE_{nm} *modes*

$$\alpha + i\Delta\beta = \frac{(1-\nu^2)^{1/2}}{a}\,\frac{n^2}{p^2 - n^2}\,(\xi + i\eta) \tag{9b}$$

It may be of interest to compare the attenuation constants given by (9) with the results obtained by calculating the power dissipated in the walls of a pipe[9] which has different resistances in the circumferential and longitudinal directions. If the wall resistance for circumferential currents is represented by R_θ and for longitudinal currents by R_z, the expressions for α are

TM_{nm} *modes*

$$\alpha = \frac{R_z}{(\mu_0/\epsilon_0)^{1/2} a(1 - \nu^2)^{1/2}}$$

TE_{nm} *modes*

$$\alpha = \frac{R_\theta \nu^2 + R_z(n/p)^2(1 - \nu^2)}{(\mu_0/\epsilon_0)^{1/2} a(1 - \nu^2)^{1/2}} \frac{p^2}{p^2 - n^2}$$

The results for ordinary metallic pipe are obtained by setting

$$R_\theta = R_z = R = (\omega\mu_0/2\sigma)^{1/2}$$

If $R_\theta = 0$, the expressions above agree with (9), inasmuch as $\xi = R(\epsilon_0/\mu_0)^{1/2}$ when the jacket conductivity is large.

$\psi = \tan^{-1} n(1 - \nu^2)^{1/2}/p\nu$, $n \neq 0$

For this value of ψ the circularly polarized TE_{nm} mode which varies as $\exp(-in\theta)$ has low attenuation. (We assume $n \neq 0$, since the case $n = 0$ has been treated above.) One of the properties of helix waveguide is the difference in propagation between right and left circularly polarized TE_{nm} modes. By properly designing the helix angle for the frequency, mode, and size of guide, the loss to one of the polarizations can be made very low. If the jacket is lossy enough the attenuation of the other polarization should be quite high. Thus only one of the circularly polarized modes should be propagated through a long pipe. Such a helix has features analogous to the optical properties of levulose and dextrose solutions, which distinguish between left and right circularly polarized light.

Let α_n be the attenuation constant of the mode which varies as $\exp(-in\theta)$, and α_{-n} the attenuation constant of the mode which varies

[9] S. A. Schelkunoff, Electromagnetic Waves, van Nostrand, New York, 1943, pp. 385–387.

as $\exp(+in\theta)$. Then from (7b), for any pitch angle ψ,

$$\alpha_{-n} = \frac{\xi}{a}\frac{p^2}{p^2 - n^2}\frac{\nu^2}{(1-\nu^2)^{1/2}}\frac{[\tan\psi + n(1-\nu^2)^{1/2}/p\nu]^2}{1+\tan^2\psi}$$

$$\alpha_n = \frac{\xi}{a}\frac{p^2}{p^2 - n^2}\frac{\nu^2}{(1-\nu^2)^{1/2}}\frac{[\tan\psi - n(1-\nu^2)^{1/2}/p\nu]^2}{1+\tan^2\psi}$$

$$\alpha_{-n} - \alpha_n = 4\frac{\xi}{a}\frac{np}{p^2 - n^2}\frac{\nu\tan\psi}{1+\tan^2\psi}$$

The mode which varies as $\exp(-in\theta)$ has lower loss if ψ and n have the same sign.

The TM_{nm} attenuation constants are independent of polarization and are given by (7a).

$\psi = 90°$

These "helices," with wires parallel to the axis of the waveguide, should propagate TM_{nm} modes with losses approximately the same as in copper pipe. For the TE_{nm} modes, (7b) gives

TE_{nm} modes

$$\alpha + i\Delta\beta = \frac{\nu^2}{a(1-\nu^2)^{1/2}}\frac{p^2}{p^2 - n^2}(\xi + i\eta)$$

IV. NUMERICAL SOLUTIONS FOR ZERO-PITCH HELICES

The main interest in helix waveguide is for small pitch angles where the TE_{01} attenuation is very low. The propagation constants of various lossy modes in helix guides of zero pitch have been calculated by solving the characteristic equation (6) numerically. These calculations will now be described.

Equation (6) is first simplified by setting $\psi = 0$ and replacing the Hankel functions with their asymptotic expressions. The condition for validity of the asymptotic expressions, namely

$$|\zeta_2 a| \gg |(4n^2 - 1)/8|$$

is well satisfied in all cases to be treated here. Equation (6) may then be rearranged in the dimensionless form

$$F_n(\zeta_1 a) = (\zeta_2 a)^3[(nha)^2 J_n^2(\zeta_1 a) - (\beta_0 a)^2(\zeta_1 a)^2 J_n'^2(\zeta_1 a)] - i(\zeta_1 a)^3[(nha)^2 + (\beta_0 a)^2(\epsilon' - i\epsilon'')(\zeta_2 a)^2]J_n'(\zeta_1 a)J_n(\zeta_1 a) \quad (10)$$
$$= 0$$

There is no difference between the propagation constants of right and

left circularly polarized waves when $\psi = 0$. Using the relationships

$$\zeta_2 a = [(\zeta_1 a)^2 + (\beta_0 a)^2 (\epsilon' - i\epsilon'' - 1)]^{1/2}, \qquad \mathrm{Im}\zeta_2 a < 0$$

$$ha = [(\beta_0 a)^2 - (\zeta_1 a)^2]^{1/2}, \qquad \mathrm{Im}\, ha < 0$$

it is clear that $F_n(\zeta_1 a)$ is an even function of $\zeta_1 a$, involving the parameters $\beta_0 a$ $(= 2\pi a/\lambda_0)$, ϵ', ϵ'', and n.

When specific values have been assigned to $\beta_0 a$, ϵ', and ϵ'', roots of (10) can be found numerically by the straightforward procedure of evaluating $F_n(\zeta_1 a)$ at a regular network of points in the plane of the complex variable $\zeta_1 a$, plotting the families of curves Re $F_n = 0$ and Im $F_n = 0$, and reading off the values of $\zeta_1 a$ corresponding to the intersections of curves of the two families.

The procedure just outlined has been applied to the cases $n = 0$ and $n = 1$. When $n = 0$ one can take out of $F_0(\zeta_1 a)$ the factor $J_0'(\zeta_1 a)$, whose roots correspond to the TE_{0m} modes; the roots of the other factor are the TM_{0m}-limit modes. When $n = 1$ the function $F_1(\zeta_1 a)$ does not factor, and its roots correspond to both TE_{1m}-limit and TM_{1m}-limit modes. If the jacket conductivity is high it is easy to identify the various limit modes, and a given mode can be traced continuously if the conductivity is decreased in sufficently small steps.

The numerical calculations were set up, more or less arbitrarily, to cover the region $0 \leqq \mathrm{Re}\ \zeta_1 a \leqq 10$, $-10 \leqq \mathrm{Im}\ \zeta_1 a \leqq 10$, for each set of parameter values. A few plots of Re F_n and Im F_n made it apparent that for propagating modes the roots in this region are all in the first quadrant and usually near the real axis. The entire process of solution was then programmed by Mrs. F. M. Laurent for automatic execution on an IBM 650 magnetic drum calculator. The calculator first evaluated $F_n(\zeta_1 a)$ at a network of points spaced half a unit apart in both directions, then examined the sign changes of Re F_n and Im F_n around each elementary square. If it appeared that a particular square might contain a root of F_n, the values of F_n at the four corner points were fitted by an interpolating cubic polynomial[10] which was then solved. If the cubic had a root inside the given square, this was recorded as an approximate root of F_n. The normalized propagation constant $iha = \alpha a + i\beta a$ was also recorded for each root.

The calculated roots $\zeta_1 a$ and the normalized propagation constants are summarized in Tables I(a) to I(f), which relate to the following cases:

Table I(a) — $\beta_0 a = 29.554$, $\epsilon' = 4$, ϵ'' variable

Table I(b) — $\beta_0 a = 29.554$, $\epsilon' = 100$, ϵ'' variable

Table I(c) — $\beta_0 a = 29.554$, $\epsilon' = \epsilon''$, both variable

[10] A. N. Lowan and H. E. Salzer, Jour. Math. and Phys., **23**, p. 157, 1944.

Table I(d) — $\beta_0 a = 12.930$, $\epsilon' = 4$, ϵ'' variable

Table I(e) — $\beta_0 a = 12.930$, $\epsilon' = \epsilon''$, both variable

Table I(f) — $\beta_0 a = 6.465$, $\epsilon' = 4$, ϵ'' variable

The three values of $\beta_0 a$ correspond to waveguides of diameter 2 inches, $\frac{7}{8}$ inch, and $\frac{7}{16}$ inch at $\lambda_0 = 5.4$ mm. The jacket materials (mostly carbon-loaded resins) which have been tested to date show a range of relative permittivities roughly from 4 to 100. There is some indication that the permittivity of a carbon-loaded resin increases as its conductivity increases; this suggested consideration of the case $\epsilon' = \epsilon''$.

The tables cover the range from $\epsilon'' = 1000$ down to $\epsilon'' = 1$ at small enough intervals so that the general course of each mode can be followed. It is worth noting that at 5.4 mm a resistivity $(1/\sigma)$ of 1 ohm cm corresponds to $\epsilon'' = 32$. Copper at this frequency has an ϵ'' of approximately 2×10^7.

In general the tables include the modes derived from $F_0(\zeta_1 a)$ whose limits are TM_{01}, TM_{02}, and TM_{03}, and the modes derived from $F_1(\zeta_1 a)$ whose limits are TE_{11}, TM_{11}, TE_{12}, TM_{12}, and TE_{13} (except that in the $\frac{7}{16}$-inch guide TM_{03}, TM_{12}, and TE_{13} are cut off). Some results are given for the TM_{13}-limit mode, namely those which satisfy the arbitrary criterion Re $\zeta_1 a \leqq 10$; but these results are incomplete because for large ϵ'' the corresponding root of $F_1(\zeta_1 a)$ approaches 10.173. Furthermore for small values of ϵ'' the attenuation constants of a few of the TM-limit modes become quite large and the corresponding values of $\zeta_1 a$ move far away from the origin. Since our object was to make a general survey rather than to investigate any particular mode exhaustively, we did not attempt to pursue these modes outside the region originally proposed for study.

The results of the IBM calculations are recorded in Table I to three decimal places. Since the roots $\zeta_1 a$ were obtained by cubic interpolation in a square of side 0.5, the last place is not entirely reliable; but spot checks on a few of the roots by successive approximations indicate that it is probably not off by more than one or two units. The propagation constants of some of the relatively low-loss modes (especially TE_{12} and TE_{13}, whose wall currents are largely circumferential) were calculated from the approximate formulas,* as noted in the tables. The attenuation

(*Text continued on page 1375*)

* The formulas used were (A9) and (A10) of the appendix, which are slightly more accurate than (7) of the text.

TABLE I(a) — 2-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 29.554$) WITH $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	γa	$\alpha a + i\beta a$
TM_{01}	∞	2.405	29.456i
	1000	2.154 + 0.384i	0.028 + 29.478i
	250	2.094 + 0.974i	0.069 + 29.496i
	100	2.408 + 1.679i	0.137 + 29.504i
	90	2.482 + 1.772i	0.149 + 29.503i
	80	2.579 + 1.878i	0.164 + 29.502i
	64	2.804 + 2.083i	0.198 + 29.495i
	40	3.519 + 2.547i	0.304 + 29.456i
	25	4.604 + 3.165i	0.496 + 29.369i
	16	5.870 + 3.763i	0.756 + 29.219i
	10	7.564 + 4.131i	1.082 + 28.887i
	8	8.464 + 4.158i	1.229 + 28.646i
TM_{02}	∞	5.520	29.034i
	1000	5.399 + 0.127i	0.024 + 29.057i
	250	5.274 + 0.268i	0.049 + 29.081i
	100	5.109 + 0.445i	0.078 + 29.113i
	90	5.081 + 0.472i	0.082 + 29.118i
	80	5.047 + 0.504i	0.087 + 29.125i
	64	4.968 + 0.569i	0.097 + 29.139i
	40	4.716 + 0.701i	0.113 + 29.184i
	25	4.375 + 0.677i	0.101 + 29.237i
	16	4.172 + 0.551i	0.079 + 29.264i
	10	4.047 + 0.448i	0.062 + 29.279i
	8	4.004 + 0.412i	0.056 + 29.285i
	4	3.905 + 0.344i	0.046 + 29.297i
	1	3.820 + 0.310i	0.040 + 29.308i
TM_{03}	∞	8.654	28.259i
	1000	8.577 + 0.078i	0.024 + 28.282i
	250	8.500 + 0.160i	0.048 + 28.306i
	100	8.408 + 0.260i	0.077 + 28.334i
	90	8.395 + 0.275i	0.081 + 28.338i
	80	8.378 + 0.293i	0.086 + 28.343i
	64	8.344 + 0.330i	0.097 + 28.354i
	40	8.253 + 0.424i	0.123 + 28.382i
	25	8.125 + 0.545i	0.156 + 28.421i
	16	7.943 + 0.678i	0.189 + 28.475i
	10	7.658 + 0.779i	0.209 + 28.556i
	8	7.511 + 0.780i	0.205 + 28.595i
	4	7.200 + 0.693i	0.174 + 28.673i
	1	6.986 + 0.612i	0.149 + 28.724i
TE_{11}	∞	1.841	29.497i
	1000	1.703 + 0.234i	0.014 + 29.506i
	250	1.764 + 0.630i	0.038 + 29.508i
	100	2.465 + 0.963i	0.081 + 29.467i
	90	2.660 + 0.748i	0.068 + 29.444i
	80	2.633 + 0.604i	0.054 + 29.443i
	64	2.594 + 0.464i	0.041 + 29.444i
	40	2.546 + 0.312i	0.027 + 29.446i
	25	2.508 + 0.226i	0.019 + 29.448i
	16	2.481 + 0.176i	0.015 + 29.450i
	10	2.455 + 0.140i	0.012 + 29.452i
	8	2.445 + 0.129i	0.011 + 29.453i
	4	2.418 + 0.106i	0.009 + 29.455i
	1	2.394 + 0.095i	0.008 + 29.457i

TABLE I(a) — *Continued*

Limit Mode	ϵ''	γ_{1a}	$\alpha_a + i\beta_a$
TM_{11}	∞	3.832	29.305i
	1000	3.652 + 0.197i	0.024 + 29.328i
	250	3.457 + 0.440i	0.052 + 29.355i
	100	2.978 + 0.880i	0.089 + 29.417i
	90	2.821 + 1.215i	0.116 + 29.445i
	80	2.945 + 1.476i	0.148 + 29.444i
	64	3.146 + 1.868i	0.200 + 29.446i
	40	3.728 + 2.564i	0.325 + 29.432i
	25	4.659 + 3.175i	0.504 + 29.361i
	16	5.921 + 3.727i	0.756 + 29.204i
	10	7.613 + 4.135i	1.090 + 28.875i
	8	8.487 + 4.153i	1.231 + 28.639i
TE_{12}	∞	5.331	29.069i
	1000		0.0008 + 29.070i*
	250		0.0016 + 29.071i*
	100		0.0026 + 29.072i*
	64		0.0033 + 29.072i*
	40		0.0042 + 29.073i*
	25		0.0055 + 29.074i*
	10		0.0092 + 29.075i*
	4	5.297 + 0.072i	0.013 + 29.076i
	1	5.322 + 0.096i	0.018 + 29.071i
TM_{12}	∞	7.016	28.710i
	1000	6.918 + 0.099i	0.024 + 28.733i
	250	6.821 + 0.203i	0.048 + 28.757i
	100	6.701 + 0.330i	0.077 + 28.786i
	90	6.683 + 0.349i	0.081 + 28.791i
	80	6.660 + 0.372i	0.086 + 28.796i
	64	6.612 + 0.419i	0.096 + 28.808i
	40	6.475 + 0.535i	0.120 + 28.841i
	25	6.253 + 0.655i	0.142 + 28.893i
	16	5.965 + 0.682i	0.141 + 28.954i
	10	5.719 + 0.590i	0.116 + 29.002i
	8	5.641 + 0.541i	0.105 + 29.016i
	4	5.471 + 0.419i	0.079 + 29.047i
	1	5.317 + 0.347i	0.063 + 29.074i
TE_{13}	∞	8.536	28.295i
	1000		0.0003 + 28.295i*
	250		0.0006 + 28.295i*
	100		0.0010 + 28.296i*
	64		0.0012 + 28.296i*
	40		0.0016 + 28.296i*
	25		0.0020 + 28.296i*
	10		0.0034 + 28.297i*
	4		0.0050 + 28.296i*
	1		0.0058 + 28.295i*
TM_{13}	∞	10.173	27.748i
	100	9.963 + 0.219i	0.078 + 27.825i
	90	9.952 + 0.231i	0.083 + 27.829i
	80	9.938 + 0.246i	0.088 + 27.834i
	64	9.911 + 0.277i	0.098 + 27.845i
	40	9.840 + 0.356i	0.126 + 27.870i
	25	9.746 + 0.460i	0.161 + 27.905i
	16	9.625 + 0.591i	0.204 + 27.950i
	10	9.433 + 0.757i	0.255 + 28.020i
	8	9.305 + 0.837i	0.278 + 28.065i
	4	8.836 + 0.898i	0.281 + 28.218i
	1	8.485 + 0.781i	0.234 + 28.322i

* Approximate formula.

TABLE I(b) — 2-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 29.554$) WITH $\epsilon' = 100$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	$\zeta_1 a$	$\alpha a + i\beta a$
TM_{01}	∞	2.405	29.456i
	1000	2.178 + 0.391i	0.029 + 29.476i
	250	2.291 + 0.885i	0.069 + 29.479i
	100	2.677 + 1.062i	0.097 + 29.452i
	80	2.764 + 1.047i	0.098 + 29.443i
	64	2.834 + 1.019i	0.098 + 29.436i
	40	2.928 + 0.950i	0.094 + 29.424i
	25	2.973 + 0.893i	0.090 + 29.418i
	10	3.004 + 0.831i	0.085 + 29.413i
	4	3.013 + 0.806i	0.083 + 29.411i
	1	3.016 + 0.793i	0.081 + 29.411i
TM_{02}	∞	5.520	29.034i
	1000	5.406 + 0.133i	0.025 + 29.056i
	250	5.339 + 0.298i	0.055 + 29.069i
	100	5.372 + 0.473i	0.087 + 29.066i
	80	5.398 + 0.508i	0.094 + 29.062i
	64	5.429 + 0.535i	0.100 + 29.056i
	40	5.492 + 0.566i	0.107 + 29.045i
	25	5.540 + 0.573i	0.109 + 29.036i
	10	5.589 + 0.569i	0.109 + 29.027i
	4	5.608 + 0.563i	0.109 + 29.023i
	1	5.617 + 0.560i	0.108 + 29.021i
TM_{03}	∞	8.654	28.259i
	1000	8.581 + 0.082i	0.025 + 28.281i
	250	8.537 + 0.179i	0.054 + 28.295i
	100	8.548 + 0.279i	0.084 + 28.292i
	80	8.561 + 0.300i	0.091 + 28.289i
	64	8.575 + 0.317i	0.096 + 28.285i
	40	8.606 + 0.339i	0.103 + 28.276i
	25	8.630 + 0.348i	0.106 + 28.268i
	10	8.658 + 0.352i	0.108 + 28.260i
	4	8.669 + 0.352i	0.108 + 28.257i
	1	8.675 + 0.351i	0.108 + 28.255i
TE_{11}	∞	1.841	29.497i
	1000	1.719 + 0.236i	0.014 + 29.505i
	250	1.871 + 0.504i	0.032 + 29.499i
	100	2.132 + 0.484i	0.035 + 29.481i
	80	2.161 + 0.451i	0.033 + 29.479i
	64	2.178 + 0.420i	0.031 + 29.477i
	40	2.191 + 0.372i	0.028 + 29.475i
	25	2.192 + 0.343i	0.026 + 29.475i
	10	2.190 + 0.316i	0.023 + 29.475i
	4	2.188 + 0.306i	0.023 + 29.475i
	1	2.187 + 0.301i	0.022 + 29.475i

TABLE I(b) — *Continued*

Limit Mode	ϵ''	$\beta_0 a$	$\alpha a + i\beta a$
TM_{11}	∞	3.832	29.305i
	1000	3.663 + 0.204i	0.026 + 29.327i
	250	3.579 + 0.485i	0.059 + 29.341i
	100	3.715 + 0.788i	0.100 + 29.331i
	80	3.787 + 0.826i	0.107 + 29.322i
	64	3.856 + 0.843i	0.111 + 29.314i
	40	3.969 + 0.836i	0.113 + 29.299i
	25	4.043 + 0.817i	0.113 + 29.288i
	10	4.100 + 0.777i	0.109 + 29.279i
	4	4.119 + 0.759i	0.107 + 29.276i
	1	4.128 + 0.749i	0.106 + 29.274i
TE_{12}	∞	5.331	29.069i
	1000		0.0008 + 29.070i*
	250		0.0018 + 29.071i*
	100		0.0028 + 29.071i*
	64		0.0032 + 29.070i*
	40		0.0034 + 29.070i*
	25		0.0035 + 29.070i*
	10		0.0036 + 29.070i*
	4		0.0036 + 29.070i*
	1		0.0036 + 29.069i*
TM_{12}	∞	7.016	28.710i
	1000	6.923 + 0.103i	0.025 + 28.732i
	250	6.868 + 0.226i	0.054 + 28.746i
	100	6.885 + 0.355i	0.085 + 28.743i
	80	6.902 + 0.381i	0.092 + 28.740i
	64	6.922 + 0.403i	0.097 + 28.735i
	40	6.965 + 0.429i	0.104 + 28.725i
	25	7.000 + 0.440i	0.107 + 28.717i
	10	7.037 + 0.443i	0.109 + 28.708i
	4	7.051 + 0.441i	0.108 + 28.704i
	1	7.058 + 0.440i	0.108 + 28.703i
TE_{13}	∞	8.536	28.295i
	1000		0.0003 + 28.295i*
	250		0.0007 + 28.295i*
	100		0.0010 + 28.295i*
	64		0.0012 + 28.295i*
	40		0.0013 + 28.295i*
	25		0.0013 + 28.295i*
	10		0.0013 + 28.295i*
	4		0.0013 + 28.295i*
	1		0.0013 + 28.295i*

* Approximate formula.

TABLE I(c) — 2-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 29.554$) WITH $\epsilon' = \epsilon''$

Limit Mode	ϵ' and ϵ''	$\zeta_1 a$	$\alpha a + i\beta a$
TM_{01}	∞	2.405	29.456i
	1000	2.338 + 0.341i	0.027 + 29.464i
	250	2.418 + 0.707i	0.058 + 29.464i
	100	2.677 + 1.062i	0.097 + 29.452i
	64	2.925 + 1.226i	0.122 + 29.435i
	40	3.309 + 1.324i	0.149 + 29.399i
	32	3.540 + 1.299i	0.156 + 29.371i
	25	3.787 + 1.162i	0.150 + 29.334i
	16	3.946 + 0.800i	0.108 + 29.301i
	12	3.950 + 0.647i	0.087 + 29.296i
	10	3.946 + 0.573i	0.077 + 29.295i
	4	3.905 + 0.344i	0.046 + 29.297i
	2	3.869 + 0.252i	0.033 + 29.301i
	1	3.820 + 0.185i	0.024 + 29.307i
TM_{02}	∞	5.520	29.034i
	1000	5.469 + 0.136i	0.026 + 29.044i
	250	5.423 + 0.282i	0.053 + 29.054i
	100	5.372 + 0.473i	0.087 + 29.066i
	64	5.337 + 0.624i	0.115 + 29.075i
	40	5.294 + 0.874i	0.159 + 29.090i
	32	5.279 + 1.061i	0.193 + 29.099i
	25	5.319 + 1.367i	0.250 + 29.105i
	16	5.852 + 1.969i	0.397 + 29.039i
	12	6.472 + 2.178i	0.487 + 28.923i
	10	7.026 + 2.198i	0.536 + 28.796i
TM_{03}	∞	8.654	28.259i
	1000	8.620 + 0.085i	0.026 + 28.269i
	250	8.587 + 0.173i	0.052 + 28.280i
	100	8.548 + 0.279i	0.084 + 28.292i
	64	8.521 + 0.355i	0.107 + 28.302i
	40	8.483 + 0.461i	0.138 + 28.315i
	32	8.458 + 0.526i	0.157 + 28.323i
	25	8.425 + 0.611i	0.182 + 28.335i
	16	8.330 + 0.824i	0.242 + 28.369i
	12	8.206 + 1.037i	0.300 + 28.413i
	10	8.034 + 1.240i	0.350 + 28.471i
	4	7.200 + 0.693i	0.174 + 28.673i
	2	7.098 + 0.483i	0.120 + 28.694i
	1	6.998 + 0.349i	0.085 + 28.716i
TE_{11}	∞	1.841	29.497i
	1000	1.810 + 0.190i	0.012 + 29.499i
	250	1.911 + 0.384i	0.025 + 29.495i
	100	2.132 + 0.484i	0.035 + 29.481i
	64	2.270 + 0.453i	0.035 + 29.470i
	40	2.365 + 0.366i	0.029 + 29.462i
	32	2.389 + 0.324i	0.026 + 29.459i
	25	2.406 + 0.281i	0.023 + 29.457i
	16	2.420 + 0.219i	0.018 + 29.456i
	12	2.424 + 0.187i	0.015 + 29.455i
	10	2.424 + 0.169i	0.014 + 29.455i
	4	2.418 + 0.106i	0.009 + 29.455i
	2	2.409 + 0.078i	0.006 + 29.456i
	1	2.394 + 0.056i	0.005 + 29.457i

TABLE I(c)—*Continued*

Limit Mode	ϵ' and ϵ''	βa	$\alpha a + i\beta a$
TM_{11}	∞	3.832	29.305i
	1000	3.759 + 0.203i	0.026 + 29.315i
	250	3.714 + 0.439i	0.056 + 29.323i
	100	3.715 + 0.788i	0.100 + 29.331i
	64	3.797 + 1.070i	0.139 + 29.329i
	40	4.080 + 1.400i	0.195 + 29.305i
	32	4.276 + 1.550i	0.226 + 29.285i
	25	4.586 + 1.661i	0.260 + 29.245i
	16	5.359 + 1.579i	0.291 + 29.109i
	12	5.587 + 1.043i	0.201 + 29.041i
	10	5.560 + 0.859i	0.164 + 29.040i
	4	5.471 + 0.419i	0.079 + 29.047i
	2	5.438 + 0.249i	0.047 + 29.051i
	1	5.444 + 0.131i	0.025 + 29.049i
TE_{12}	∞	5.331	29.069i
	1000		0.0009 + 29.070i*
	250		0.0018 + 29.070i*
	100		0.0028 + 29.071i*
	64		0.0035 + 29.071i*
	40		0.0044 + 29.071i*
	25		0.0055 + 29.072i*
	10		0.0087 + 29.073i*
	4	5.297 + 0.072i	0.013 + 29.076i
	2	5.272 + 0.108i	0.020 + 29.080i
	1	5.198 + 0.132i	0.023 + 29.094i
TM_{12}	∞	7.016	28.710i
	1000	6.971 + 0.107i	0.026 + 28.721i
	250	6.931 + 0.217i	0.052 + 28.731i
	100	6.885 + 0.355i	0.085 + 28.743i
	64	6.852 + 0.457i	0.109 + 28.753i
	40	6.801 + 0.610i	0.144 + 28.768i
	32	6.768 + 0.708i	0.167 + 28.778i
	25	6.720 + 0.850i	0.198 + 28.793i
	16	6.562 + 1.359i	0.309 + 28.850i
	12	6.869 + 2.095i	0.499 + 28.825i
	10	7.322 + 2.374i	0.605 + 28.737i
TE_{13}	∞	8.536	28.295i
	1000		0.0003 + 28.295i*
	250		0.0007 + 28.295i*
	100		0.0010 + 28.295i*
	64		0.0013 + 28.295i*
	40		0.0016 + 28.295i*
	25		0.0021 + 28.295i*
	10		0.0032 + 28.296i*
	4		0.0050 + 28.296i*
	1		0.0094 + 28.295i*
TM_{13}	∞	10.173	27.748i
	25	9.981 + 0.497i	0.178 + 27.823i
	16	9.910 + 0.652i	0.232 + 27.852i
	12	9.841 + 0.785i	0.277 + 27.880i
	10	9.776 + 0.893i	0.313 + 27.907i
	4	8.836 + 0.898i	0.281 + 28.218i
	2	8.656 + 0.596i	0.183 + 28.265i
	1	8.523 + 0.409i	0.123 + 28.302i

* Approximate formula.

TABLE I(d)—$\frac{7}{8}$-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 12.930$) WITH $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	βa	$\alpha a + i\beta a$
TM_{01}	∞	2.405	12.704i
	1000	2.286 + 0.140i	0.025 + 12.727i
	250	2.183 + 0.324i	0.056 + 12.749i
	100	2.113 + 0.595i	0.098 + 12.771i
	64	2.114 + 0.800i	0.132 + 12.782i
	40	2.185 + 1.072i	0.183 + 12.790i
	25	2.377 + 1.369i	0.255 + 12.786i
	10	3.212 + 1.699i	0.431 + 12.647i
	6.4	3.694 + 1.440i	0.426 + 12.482i
	4.0	3.765 + 1.029i	0.312 + 12.416i
	2.5	3.700 + 0.853i	0.254 + 12.421i
	1.0	3.624 + 0.733i	0.214 + 12.435i
TM_{02}	∞	5.520	11.692i
	1000	5.468 + 0.054i	0.025 + 11.717i
	250	5.416 + 0.111i	0.051 + 11.742i
	100	5.356 + 0.183i	0.083 + 11.770i
	64	5.317 + 0.235i	0.106 + 11.789i
	40	5.266 + 0.308i	0.137 + 11.814i
	25	5.206 + 0.410i	0.180 + 11.844i
	10	5.073 + 0.772i	0.328 + 11.923i
	6.4	5.095 + 1.137i	0.485 + 11.948i
	4.0	5.486 + 1.429i	0.664 + 11.814i
	2.5	5.818 + 1.379i	0.689 + 11.650i
	1.0	6.041 + 1.188i	0.624 + 11.511i
TM_{03}	∞	8.654	9.607i
	1000	8.620 + 0.034i	0.030 + 9.637i
	250	8.587 + 0.069i	0.061 + 9.667i
	100	8.550 + 0.111i	0.098 + 9.701i
	64	8.525 + 0.141i	0.124 + 9.723i
	40	8.494 + 0.183i	0.160 + 9.752i
	25	8.459 + 0.239i	0.207 + 9.785i
	10	8.393 + 0.411i	0.350 + 9.851i
	6.4	8.386 + 0.532i	0.452 + 9.866i
	4.0	8.426 + 0.668i	0.571 + 9.847i
	2.5	8.515 + 0.769i	0.669 + 9.784i
	1.0	8.676 + 0.824i	0.741 + 9.651i
TE_{11}	∞	1.841	12.798i
	1000	1.767 + 0.074i	0.010 + 12.809i
	250	1.717 + 0.191i	0.026 + 12.817i
	100	1.706 + 0.368i	0.049 + 12.822i
	64	1.734 + 0.500i	0.068 + 12.823i
	40	1.857 + 0.656i	0.095 + 12.813i
	25	2.126 + 0.773i	0.129 + 12.778i
	10	2.436 + 0.411i	0.079 + 12.706i
	6.4	2.413 + 0.316i	0.060 + 12.707i
	4.0	2.386 + 0.262i	0.049 + 12.711i
	2.5	2.364 + 0.234i	0.043 + 12.714i
	1.0	2.341 + 0.212i	0.039 + 12.718i

TABLE I(d) — *Continued*

Limit Mode	ϵ''	$\beta_1 a$	$\alpha a + i\beta a$
TM_{11}	∞	3.832	12.349i
	1000	3.750 + 0.081i	0.025 + 12.375i
	250	3.676 + 0.171i	0.051 + 12.398i
	100	3.588 + 0.290i	0.084 + 12.426i
	64	3.530 + 0.382i	0.108 + 12.445i
	40	3.447 + 0.516i	0.143 + 12.474i
	25	3.329 + 0.757i	0.201 + 12.519i
	10	3.749 + 1.664i	0.499 + 12.496i
	6.4	4.275 + 1.750i	0.606 + 12.343i
	4.0	4.701 + 1.553i	0.600 + 12.160i
	2.5	4.843 + 1.274i	0.511 + 12.067i
	1.0	4.844 + 1.031i	0.415 + 12.040i
TE_{12}	∞	5.331	11.780i
	1000		0.0007 + 11.780i*
	250		0.0015 + 11.781i*
	100		0.0024 + 11.782i*
	64		0.0030 + 11.782i*
	40		0.0039 + 11.783i*
	25		0.0051 + 11.784i*
	10		0.0085 + 11.785i*
	4		0.0125 + 11.784i*
	1		0.0146 + 11.781i*
TM_{12}	∞	7.016	10.861i
	1000	6.972 + 0.043i	0.027 + 10.889i
	250	6.930 + 0.087i	0.055 + 10.917i
	100	6.883 + 0.141i	0.088 + 10.947i
	64	6.853 + 0.179i	0.112 + 10.967i
	40	6.814 + 0.233i	0.144 + 10.992i
	25	6.769 + 0.305i	0.187 + 11.023i
	10	6.679 + 0.541i	0.326 + 11.090i
	6.4	6.670 + 0.718i	0.431 + 11.109i
	4.0	6.755 + 0.935i	0.570 + 11.080i
	2.5	6.942 + 1.061i	0.671 + 10.981i
	1.0	7.193 + 1.054i	0.700 + 10.819i
TE_{13}	∞	8.536	9.712i
	1000		0.0002 + 9.712i*
	250		0.0005 + 9.712i*
	100		0.0008 + 9.712i*
	64		0.0010 + 9.713i*
	40		0.0012 + 9.713i*
	25		0.0016 + 9.713i*
	10		0.0027 + 9.713i*
	4		0.0040 + 9.713i*
	1		0.0048 + 9.712i*
TM_{13}	∞	10.173	7.980i
	10	9.949 + 0.340i	0.409 + 8.276i
	6.4	9.943 + 0.436i	0.523 + 8.293i
	4.0	9.970 + 0.543i	0.655 + 8.277i

* Approximate formula.

TABLE I(e) — $\frac{7}{8}$-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 12.930$) WITH $\epsilon' = \epsilon''$

Limit Mode	ϵ' and ϵ''	βa	$\alpha a + i\beta a$
TM_{01}	∞	2.405	12.704i
	1000	2.360 + 0.141i	0.026 + 12.714i
	250	2.339 + 0.295i	0.054 + 12.720i
	100	2.351 + 0.482i	0.089 + 12.724i
	64	2.382 + 0.608i	0.114 + 12.724i
	40	2.450 + 0.766i	0.148 + 12.720i
	25	2.573 + 0.942i	0.191 + 12.708i
	10	3.052 + 1.244i	0.301 + 12.630i
	4	3.765 + 1.029i	0.312 + 12.416i
	2	3.841 + 0.653i	0.203 + 12.366i
	1	3.768 + 0.438i	0.133 + 12.378i
TM_{02}	∞	5.520	11.692i
	1000	5.497 + 0.058i	0.027 + 11.704i
	250	5.475 + 0.118i	0.055 + 11.715i
	100	5.451 + 0.190i	0.088 + 11.727i
	64	5.435 + 0.241i	0.111 + 11.735i
	40	5.416 + 0.310i	0.143 + 11.746i
	25	5.393 + 0.402i	0.184 + 11.760i
	10	5.338 + 0.701i	0.317 + 11.802i
	4	5.486 + 1.429i	0.664 + 11.814i
	2	6.389 + 1.780i	0.996 + 11.425i
	1	6.901 + 1.040i	0.652 + 11.003i
TM_{03}	∞	8.654	9.607i
	1000	8.639 + 0.037i	0.033 + 9.621i
	250	8.624 + 0.074i	0.067 + 9.635i
	100	8.607 + 0.118i	0.105 + 9.650i
	64	8.596 + 0.148i	0.132 + 9.661i
	40	8.581 + 0.189i	0.168 + 9.675i
	25	8.563 + 0.241i	0.213 + 9.694i
	10	8.512 + 0.393i	0.344 + 9.747i
	4	8.426 + 0.668i	0.571 + 9.847i
	2	8.320 + 1.094i	0.910 + 9.999i
	1	8.812 + 1.915i	1.721 + 9.806i
TE_{11}	∞	1.841	12.798i
	1000	1.810 + 0.072i	0.010 + 12.803i
	250	1.807 + 0.161i	0.023 + 12.804i
	100	1.833 + 0.265i	0.038 + 12.802i
	64	1.870 + 0.330i	0.048 + 12.799i
	40	1.939 + 0.401i	0.061 + 12.790i
	25	2.047 + 0.459i	0.074 + 12.776i
	10	2.295 + 0.414i	0.075 + 12.732i
	4	2.386 + 0.262i	0.049 + 12.711i
	2	2.389 + 0.186i	0.035 + 12.709i
	1	2.369 + 0.129i	0.024 + 12.712i

TABLE I(e) — *Continued*

Limit Mode	ϵ' and ϵ''	γ_{1a}	$\alpha a + i\beta a$
TM_{11}	∞	3.832	12.349i
	1000	3.794 + 0.086i	0.026 + 12.361i
	250	3.766 + 0.176i	0.054 + 12.371i
	100	3.739 + 0.288i	0.087 + 12.381i
	64	3.725 + 0.369i	0.111 + 12.388i
	40	3.711 + 0.485i	0.145 + 12.396i
	25	3.708 + 0.651i	0.195 + 12.406i
	10	3.893 + 1.161i	0.365 + 12.390i
	4	4.701 + 1.553i	0.600 + 12.160i
	2	5.319 + 1.062i	0.477 + 11.843i
	1	5.241 + 0.614i	0.272 + 11.840i
TE_{12}	∞	5.331	11.780i
	1000		0.0008 + 11.780i*
	250		0.0016 + 11.780i*
	100		0.0026 + 11.781i*
	64		0.0032 + 11.781i*
	40		0.0041 + 11.781i*
	25		0.0051 + 11.782i*
	10		0.0081 + 11.783i*
	4		0.0125 + 11.784i*
	1		0.0236 + 11.782i*
TM_{12}	∞	7.016	10.861i
	1000	6.996 + 0.047i	0.030 + 10.874i
	250	6.976 + 0.094i	0.060 + 10.887i
	100	6.955 + 0.149i	0.095 + 10.902i
	64	6.942 + 0.187i	0.119 + 10.911i
	40	6.924 + 0.238i	0.151 + 10.923i
	25	6.903 + 0.305i	0.192 + 10.939i
	10	6.841 + 0.509i	0.317 + 10.988i
	4	6.755 + 0.935i	0.570 + 11.080i
	2	7.053 + 1.730i	1.106 + 11.030i
	1	8.138 + 1.672i	1.325 + 10.272i
TE_{13}	∞	8.536	9.712i
	1000		0.0003 + 9.712i*
	250		0.0005 + 9.712i*
	100		0.0008 + 9.712i*
	64		0.0010 + 9.712i*
	40		0.0013 + 9.712i*
	25		0.0016 + 9.712i*
	10		0.0025 + 9.713i*
	4		0.0040 + 9.713i*
	1		0.0076 + 9.713i*
TM_{13}	∞	10.173	7.980i
	4	9.970 + 0.543i	0.655 + 8.277i
	2	9.863 + 0.826i	0.963 + 8.457i
	1	9.698 + 1.418i	1.561 + 8.808i

* Approximate formula.

TABLE I(f)—$\frac{7}{16}$-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 6.465$) WITH $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	βa	$\alpha a + i\beta a$
TM_{01}	∞	2.405	6.001i
	1000		0.024 + 6.025i*
	250	2.287 + 0.141i	0.053 + 6.049i
	100	2.228 + 0.244i	0.090 + 6.074i
	64	2.197 + 0.324i	0.117 + 6.090i
	40	2.170 + 0.439i	0.156 + 6.107i
	25	2.169 + 0.594i	0.210 + 6.123i
	10	2.355 + 0.943i	0.364 + 6.105i
	4	2.740 + 1.040i	0.478 + 5.966i
	1	2.961 + 0.878i	0.446 + 5.830i
TM_{02}	∞	5.520	3.365i
	1000		0.043 + 3.408i*
	250	5.468 + 0.054i	0.086 + 3.450i
	100	5.439 + 0.088i	0.137 + 3.499i
	64	5.420 + 0.112i	0.172 + 3.530i
	40	5.396 + 0.146i	0.221 + 3.570i
	25	5.370 + 0.191i	0.284 + 3.616i
	10	5.327 + 0.328i	0.471 + 3.707i
	4	5.369 + 0.512i	0.740 + 3.712i
	1	5.539 + 0.614i	0.965 + 3.524i
TE_{11}	∞	1.841	6.197i
	1000		0.009 + 6.206i*
	250	1.772 + 0.069i	0.020 + 6.218i
	100	1.744 + 0.129i	0.036 + 6.227i
	64	1.731 + 0.176i	0.049 + 6.231i
	40	1.726 + 0.244i	0.068 + 6.235i
	25	1.744 + 0.334i	0.093 + 6.235i
	10	1.925 + 0.493i	0.153 + 6.193i
	4	2.121 + 0.425i	0.147 + 6.123i
	1	2.152 + 0.319i	0.112 + 6.106i
TM_{11}	∞	3.832	5.207i
	1000		0.028 + 5.235i*
	250	3.751 + 0.082i	0.058 + 5.266i
	100	3.710 + 0.134i	0.094 + 5.297i
	64	3.683 + 0.173i	0.119 + 5.317i
	40	3.650 + 0.227i	0.155 + 5.343i
	25	3.615 + 0.303i	0.204 + 5.372i
	10	3.581 + 0.546i	0.360 + 5.422i
	4	3.763 + 0.810i	0.570 + 5.350i
	1	4.038 + 0.816i	0.639 + 5.154i
TE_{12}	∞	5.331	3.657i
	1000		0.0005 + 3.657i*
	250		0.0009 + 3.658i*
	100		0.0015 + 3.658i*
	64		0.0019 + 3.659i*
	40		0.0024 + 3.659i*
	25		0.0031 + 3.659i*
	10		0.0052 + 3.660i*
	4		0.0079 + 3.660i*
	1		0.0097 + 3.658i*

* Approximate formula.

TABLE I(e) — *Continued*

Limit Mode	ϵ' and ϵ''	$\zeta_1 a$	$\alpha a + i\beta a$
TM_{11}	∞	3.832	12.349i
	1000	3.794 + 0.086i	0.026 + 12.361i
	250	3.766 + 0.176i	0.054 + 12.371i
	100	3.739 + 0.288i	0.087 + 12.381i
	64	3.725 + 0.369i	0.111 + 12.388i
	40	3.711 + 0.485i	0.145 + 12.396i
	25	3.708 + 0.651i	0.195 + 12.406i
	10	3.893 + 1.161i	0.365 + 12.390i
	4	4.701 + 1.553i	0.600 + 12.160i
	2	5.319 + 1.062i	0.477 + 11.843i
	1	5.241 + 0.614i	0.272 + 11.840i
TE_{12}	∞	5.331	11.780i
	1000		0.0008 + 11.780i*
	250		0.0016 + 11.780i*
	100		0.0026 + 11.781i*
	64		0.0032 + 11.781i*
	40		0.0041 + 11.781i*
	25		0.0051 + 11.782i*
	10		0.0081 + 11.783i*
	4		0.0125 + 11.784i*
	1		0.0236 + 11.782i*
TM_{12}	∞	7.016	10.861i
	1000	6.996 + 0.047i	0.030 + 10.874i
	250	6.976 + 0.094i	0.060 + 10.887i
	100	6.955 + 0.149i	0.095 + 10.902i
	64	6.942 + 0.187i	0.119 + 10.911i
	40	6.924 + 0.238i	0.151 + 10.923i
	25	6.903 + 0.305i	0.192 + 10.939i
	10	6.841 + 0.509i	0.317 + 10.988i
	4	6.755 + 0.935i	0.570 + 11.080i
	2	7.053 + 1.730i	1.106 + 11.030i
	1	8.138 + 1.672i	1.325 + 10.272i
TE_{13}	∞	8.536	9.712i
	1000		0.0003 + 9.712i*
	250		0.0005 + 9.712i*
	100		0.0008 + 9.712i*
	64		0.0010 + 9.712i*
	40		0.0013 + 9.712i*
	25		0.0016 + 9.712i*
	10		0.0025 + 9.713i*
	4		0.0040 + 9.713i*
	1		0.0076 + 9.713i*
TM_{13}	∞	10.173	7.980i
	4	9.970 + 0.543i	0.655 + 8.277i
	2	9.863 + 0.826i	0.963 + 8.457i
	1	9.698 + 1.418i	1.561 + 8.808i

* Approximate formula.

TABLE I(f)—$\frac{7}{16}$-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 6.465$) WITH $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	βa	$\alpha a + i\beta a$
TM_{01}	∞	2.405	6.001i
	1000		0.024 + 6.025i*
	250	2.287 + 0.141i	0.053 + 6.049i
	100	2.228 + 0.244i	0.090 + 6.074i
	64	2.197 + 0.324i	0.117 + 6.090i
	40	2.170 + 0.439i	0.156 + 6.107i
	25	2.169 + 0.594i	0.210 + 6.123i
	10	2.355 + 0.943i	0.364 + 6.105i
	4	2.740 + 1.040i	0.478 + 5.966i
	1	2.961 + 0.878i	0.446 + 5.830i
TM_{02}	∞	5.520	3.365i
	1000		0.043 + 3.408i*
	250	5.468 + 0.054i	0.086 + 3.450i
	100	5.439 + 0.088i	0.137 + 3.499i
	64	5.420 + 0.112i	0.172 + 3.530i
	40	5.396 + 0.146i	0.221 + 3.570i
	25	5.370 + 0.191i	0.284 + 3.616i
	10	5.327 + 0.328i	0.471 + 3.707i
	4	5.369 + 0.512i	0.740 + 3.712i
	1	5.539 + 0.614i	0.965 + 3.524i
TE_{11}	∞	1.841	6.197i
	1000		0.009 + 6.206i*
	250	1.772 + 0.069i	0.020 + 6.218i
	100	1.744 + 0.129i	0.036 + 6.227i
	64	1.731 + 0.176i	0.049 + 6.231i
	40	1.726 + 0.244i	0.068 + 6.235i
	25	1.744 + 0.334i	0.093 + 6.235i
	10	1.925 + 0.493i	0.153 + 6.193i
	4	2.121 + 0.425i	0.147 + 6.123i
	1	2.152 + 0.319i	0.112 + 6.106i
TM_{11}	∞	3.832	5.207i
	1000		0.028 + 5.235i*
	250	3.751 + 0.082i	0.058 + 5.266i
	100	3.710 + 0.134i	0.094 + 5.297i
	64	3.683 + 0.173i	0.119 + 5.317i
	40	3.650 + 0.227i	0.155 + 5.343i
	25	3.615 + 0.303i	0.204 + 5.372i
	10	3.581 + 0.546i	0.360 + 5.422i
	4	3.763 + 0.810i	0.570 + 5.350i
	1	4.038 + 0.816i	0.639 + 5.154i
TE_{12}	∞	5.331	3.657i
	1000		0.0005 + 3.657i*
	250		0.0009 + 3.658i*
	100		0.0015 + 3.658i*
	64		0.0019 + 3.659i*
	40		0.0024 + 3.659i*
	25		0.0031 + 3.659i*
	10		0.0052 + 3.660i*
	4		0.0079 + 3.660i*
	1		0.0097 + 3.658i*

* Approximate formula.

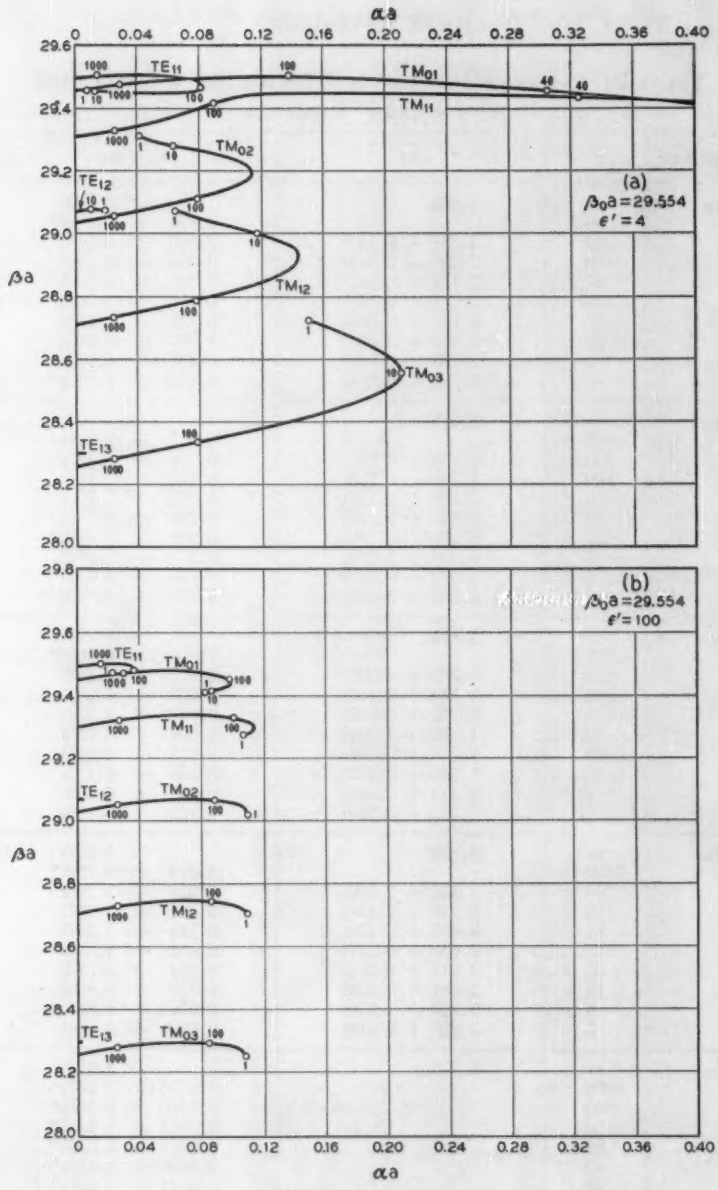


Fig. 2(a) and (b)

Fig. 2 — Plots of phase constant versus attenuation constant for modes in various helix waveguides. Representative values of ϵ'' are shown on the curves.

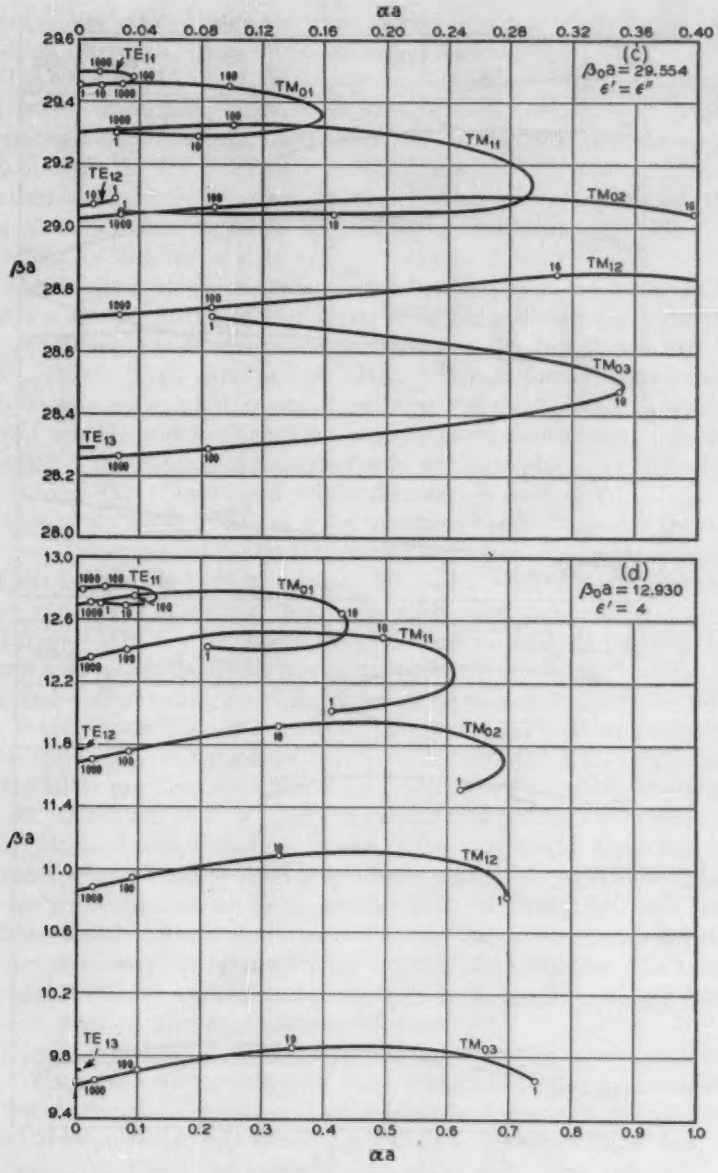

αa
βa
(c)
β0a = 29.554
ε′ = ε″
TE11
TM01
TM11
TE12
TM02
TM12
TM03
TE13
(d)
β0a = 12.930
ε′ = 4

Fig. 2(c) and (d)

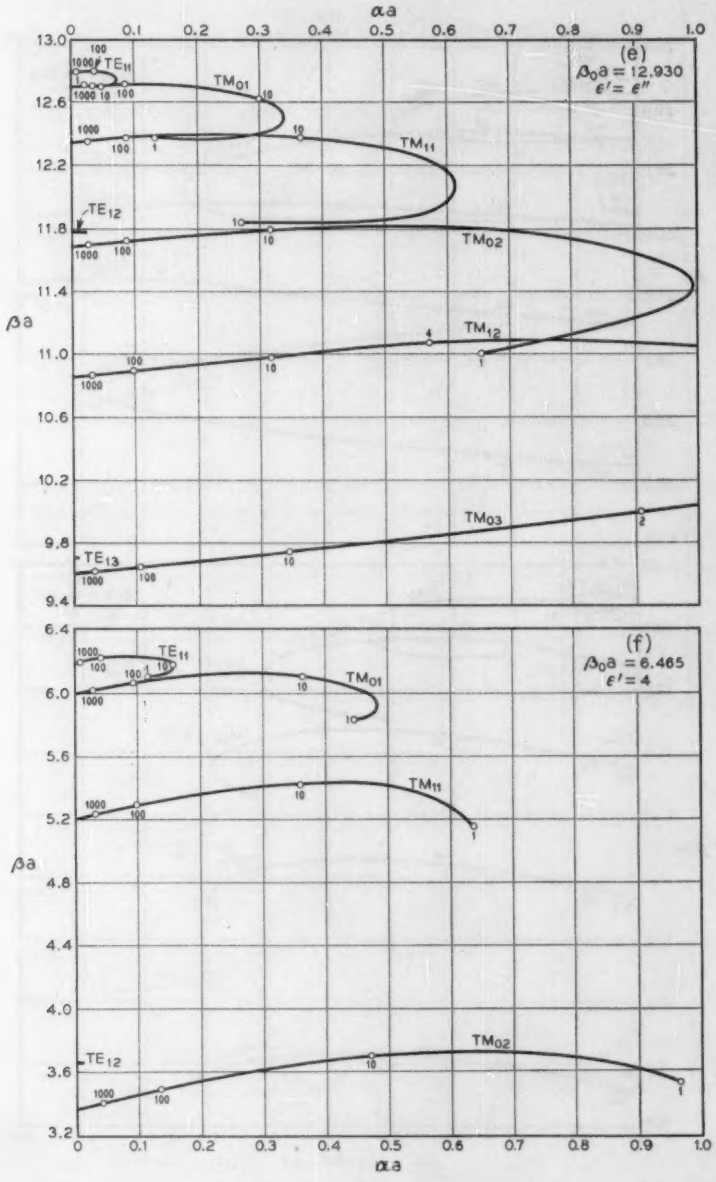

αa
βa
(e)
β0a = 12.930
ε' = ε''
TE11
TM01
TM11
TE12
TM02
TM12
TM03
TE13
(f)
β0a = 6.465
ε' = 4
TE11
TM01
TM11
TM02
TE12

Fig. 2(e) and (f)

constants calculated from the approximate formulas are given to four decimal places, i.e., usually two significant figures.

The contents of Table I are displayed graphically in Figs. 2(a) through (f), which show plots of βa vs αa for all modes except TM_{13}. Representative values of ϵ'' are indicated on the curves. Note that the scales are different for the different guide sizes, and that the βa-scale is compressed in all cases. If αa and βa were plotted on the same scale, the curves would make an initial angle of 45° with the αa-axis when $\epsilon' =$ constant, or 22.5° when $\epsilon' = \epsilon''$.

Figs. 3(a) to (f) show the normalized attenuation constants αa of various modes plotted against ϵ'' on a log-log scale. In Fig. 3(b) the curves for all TM modes would be similar to the two shown, and in Fig. 3(d) the TM_{03} curve is like TM_{12}. Although for some modes the attenuation constant increases steadily as the conductivity decreases over the range of our calculations, in many cases the attenuation passes through a maximum and then decreases as the conductivity is further decreased. This phenomenon will be discussed in Section V.

It may be noticed that in some instances the limit modes are not unique. For example, Tables I(a), with $\epsilon' = 4$, and I(c), with $\epsilon' = \epsilon''$, for the large guide have in common the case $\epsilon' = 4$, $\epsilon'' = 4$. For this case consider the circular magnetic mode corresponding to $\zeta_1 a = 3.905 + 0.344i$. If ϵ' is constant ($= 4$) while ϵ'' tends to infinity, this mode approaches the TM_{02} mode in a perfectly conducting guide; but if ϵ' and ϵ'' tend to infinity while remaining equal to each other, the same mode approaches TM_{01} in a perfectly conducting guide. Presumably the TM_{01}-limit mode in the former case coincides with the TM_{02}-limit mode in the latter case; but the value of $\zeta_1 a$ for this mode is outside the range of our calculations at $\epsilon' = \epsilon'' = 4$. A similar interchange occurs between the TM_{11}-limit and TM_{12}-limit modes in the large guide, depending on whether ϵ' is constant or ϵ' tends to infinity with ϵ''. There is no evidence of any such phenomenon in the smaller guide of Tables I(d) and I(e); but the fact that it can occur means that the limit-mode designations of modes in a lossy waveguide are not entirely unambiguous. The phenomenon is not due to the presence of the helix, since a helix of zero pitch has no effect on circular magnetic modes.

Finally it is of interest to compare the propagation constants given by the approximate formula with those obtained by numerical solution of the characteristic equation. A reasonably typical case is provided by the TM_{02}-limit mode in a 2-inch guide at $\lambda_0 = 5.4$ mm with $\epsilon' = 4$, as in Table I(a). Exact and approximate results for βa vs αa and αa vs ϵ'' are plotted in Fig. 4. As the conductivity decreases, the attenuation con-

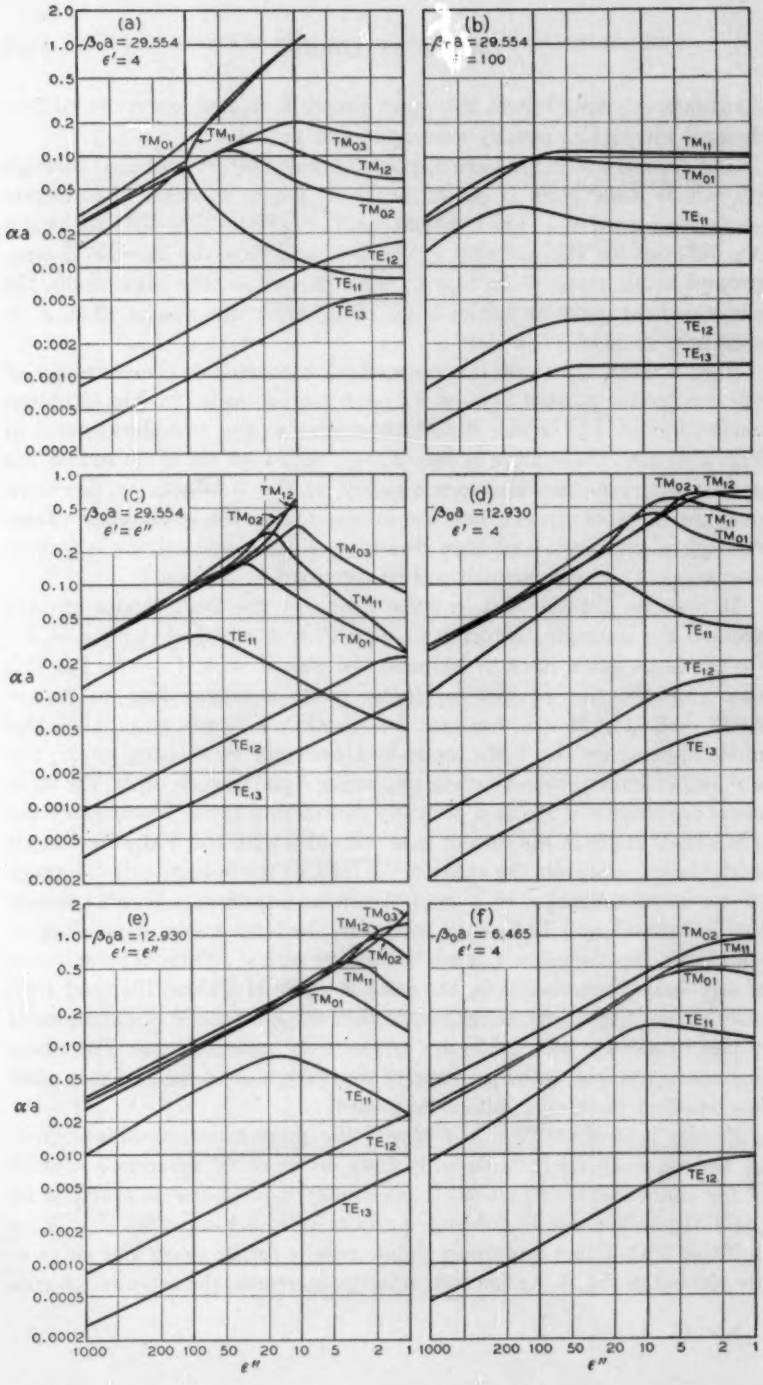


Fig. 3 — Attenuation constant as a function of jacket conductivity for modes in various helix waveguides.

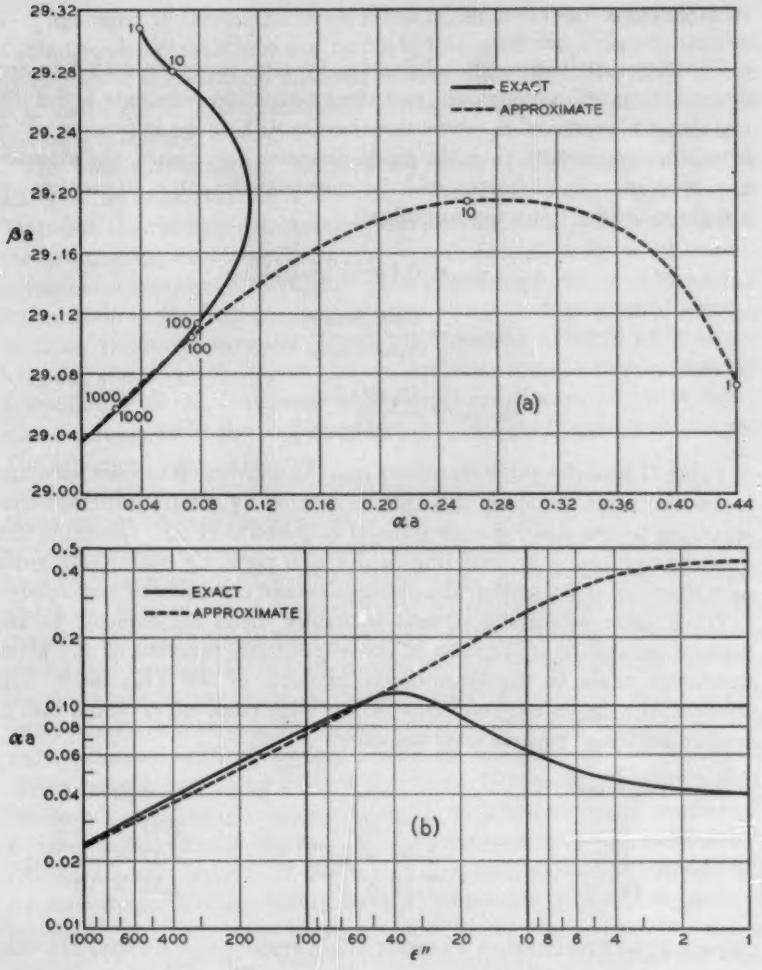


Fig. 4 — Comparison of exact and approximate formulas for the propagation constant of a typical mode (TM_{02}- limit in a guide with $\beta_0 a = 29.554$ and $\epsilon' = 4$).

stant first becomes larger, in all cases, than predicted by the approximate formula. For still lower conductivities the attenuation constant may pass through a maximum, as in the present example, and decrease again. The existence of a maximum in the attenuation vs conductivity curve is not indicated by the approximate formula.

V. DISCUSSION OF RESULTS

The dimensionless results of Section IV may easily be scaled to any desired operating wavelength, and the attenuation constants and guide wavelengths expressed in conventional units. If λ_0 is the free-space wavelength in centimeters, then the guide diameter d in inches, the attenuation constant α in db/meter, and the guide wavelength λ_g in centimeters are given by the following formulas:

$$d_{\text{in}} = 0.12532\ (\beta_0 a)(\lambda_0)_{\text{cm}}$$

$$\alpha_{\text{db/m}} = \frac{5457.5\ (\alpha a)}{(\beta_0 a)(\lambda_0)_{\text{cm}}}$$

$$(\lambda_g)_{\text{cm}} = \frac{(\beta_0 a)(\lambda_0)_{\text{cm}}}{(\beta a)}$$

Table II lists the guide diameters and the conversion factors for α and λ_g for the three values of $\beta_0 a$ used in Section IV, at frequencies corresponding to free-space wavelengths of 3.33 and 0.54 cm. The table also lists the number of propagating modes in a perfectly conducting guide as a function of $\beta_0 a$ (different polarizations are not counted separately).

When helix waveguide is used to reduce mode conversions, an important parameter is the ratio of the attenuation constant of any given unwanted mode to the attenuation constant of the TE_{01} mode. The theoretical attenuation constants of the TE_{01} mode at $\lambda_0 = 5.4$ mm in copper guides of various sizes are listed below:

Diameter	αa	αdb/m
2″	2.77×10^{-6}	9.47×10^{-4}
$\frac{7}{8}$″	1.50×10^{-5}	1.17×10^{-2}
$\frac{7}{16}$″	7.11×10^{-5}	1.11×10^{-1}

TABLE II — CONVERSION FACTORS FOR ATTENUATION CONSTANTS AND GUIDE WAVELENGTHS IN VARIOUS WAVEGUIDES

$\beta_0 a$	Propagating modes	$\lambda_0 = 3.33$ cm			$\lambda_0 = 0.54$ cm		
		Diameter (inches)	α db/meter	λ_g cm	Diameter (inches)	α db/meter	λ_g cm
29.554	227	12.33	55.5 αa	98.41/βa	2.000	342 αa	15.959/βa
12.930	44	5.40	127 αa	43.06/βa	0.875	782 αa	6.982/βa
6.465	12	2.70	253 αa	21.53/βa	0.4375	1563 αa	3.491/βa

Referring to the values of αa listed in Table I, we see that the unwanted mode attenuations can be made to exceed the TE_{01} attenuation by factors of from several hundred to several hundred thousand in the large helix guide. The attenuation ratios are somewhat smaller in the smaller guide sizes.

The attenuation versus conductivity plots of Fig. 3 show that for many of the modes there is a value of jacket conductivity, depending on the mode, the value of $\beta_0 a$, and the jacket permittivity, which maximizes the attenuation constant. Since one is accustomed to think of the attenuation constant of a waveguide as an increasing function of frequency for all sufficiently high frequencies (except for circular electric waves), or as an increasing function of wall resistance, it is worth while to see why one should really expect the attenuation constant to pass through a maximum as the frequency is increased indefinitely in an ordinary metallic guide, or as the wall resistance is increased at a fixed frequency. The argument runs as follows:

Guided waves inside a cylindrical pipe may be expressed as bundles of plane waves repeatedly reflected from the cylindrical boundary.[11] The angle which the wave normals make with the guide axis decreases as the frequency increases farther above cutoff; and the complementary angle, which is the angle of incidence of the waves upon the boundary, approaches 90°. If the walls are imperfectly conducting, the guided wave is attenuated because the reflection coefficient of the component waves at the boundary is less than unity. The theory of reflection at an imperfectly conducting surface shows that the reflection coefficient of a plane wave polarized with its electric vector in the plane of incidence first decreases with increasing angle of incidence, then passes through a deep minimum, and finally increases to unity at strictly grazing incidence.[12] For a metallic reflector, the angle of incidence corresponding to minimum reflection is very near 90°. Inasmuch as all modes in circular guide except for the circular electric family have a component of $\vec{E}$ in the plane of incidence (the plane θ = constant), one would expect the attenuation constant of each mode to pass through a maximum at a sufficiently high frequency. For example, the TM_{01} mode in a 2-inch copper guide should have maximum attenuation at a free-space wavelength in the neighborhood of 0.1 mm (100 microns), assuming the dc value for the conductivity of copper. To find the actual maximum, of course, would require the solution of a transcendental equation as in Section IV.

The circular electric waves all have $\vec{E}$ normal to the plane of incidence.

[11] Reference 9, pp. 411–412.
[12] Reference 7, pp. 507–509.

For this polarization the reflection coefficient increases steadily from its value at normal incidence to unity at grazing incidence. Thus one has an optical interpretation of the anomalous attenuation-frequency behavior of circular electric waves.

If instead of varying the frequency one imagines the wall resistance varied at a fixed frequency, he can easily convince himself that there usually exists a finite value of resistance which maximizes the attenuation constant of a given mode. An idealized illustrative example has been worked out by Schelkunoff.[13] He considers the propagation of transverse magnetic waves between parallel resistance sheets, and shows that if the sheets are far enough apart the attenuation constant increases from zero to a maximum and then falls again to zero, as the wall resistance is made to increase from zero to infinity. It may be instructive to consider that maximum power is dissipated in the lossy walls when their impedance is matched as well as possible to the wave impedance, looking normal to the walls, of the fields inside the guide.

In conclusion we mention a couple of theoretical questions which are suggested by the numerical results of Section IV.

(1) Limit modes. It has been seen that the limit which a given lossy mode approaches as the jacket conductivity becomes infinite may not be unique. Can rules be given for determining limit modes when the manner in which $| \epsilon' - i\epsilon'' |$ approaches infinity is specified?

(2) Behavior of modes as $\sigma \to 0$. It is known[14] that the number of true guided waves (i.e., exponentially propagating waves whose fields vanish at large radial distances from the guide axis) possible in a cylindrical waveguide is finite if the conductivity of the exterior medium is finite. The number is enormously large if the exterior medium is a metal; but the modes presumably disappear one by one as the conductivity is decreased. If the conductivity of the exterior medium is low enough and if its permittivity is not less than the permittivity of the interior medium, no true guided waves can exist. At what values of conductivity do the first few modes appear in a guide of given size, and how do their propagation constants behave at very low conductivities?

The complete theory of lossy-wall waveguide would appear to present quite a challenge to the applied mathematician. Fortunately the engineering usefulness of helix waveguide does not depend upon getting immediate answers to such difficult analytical questions.

---

[13] Reference 9, pp. 484-489.

[14] G. M. Roe, The Theory of Acoustic and Electromagnetic Wave Guides and Cavity Resonators, Ph.D. thesis, U. of Minn., 1947, Section 2.

APPENDIX

APPROXIMATE SOLUTION OF THE CHARACTERISTIC EQUATION

The characteristic equation (6) of the helix guide may be written in the dimensionless form

$$\left(\zeta_1 a \tan\psi - \frac{nha}{\zeta_1 a}\right)^2 \frac{J_n(\zeta_1 a)}{J_n'(\zeta_1 a)} - (\beta_0 a)^2 \frac{J_n'(\zeta_1 a)}{J_n(\zeta_1 a)}$$
$$= \frac{\zeta_1 a}{\zeta_2 a}\left[\left(\zeta_2 a \tan\psi - \frac{nha}{\zeta_2 a}\right)^2 \frac{H_n^{(2)}(\zeta_2 a)}{H_n^{(2)\prime}(\zeta_2 a)} - (\beta_0 a)^2(\epsilon' - i\epsilon'') \frac{H_n^{(2)\prime}(\zeta_2 a)}{H_n^{(2)}(\zeta_2 a)}\right] \tag{A1}$$

If $|\epsilon' - \epsilon''|$ is sufficiently large, the right side of the equation is large and either $J_n(\zeta_1 a)$ or $J_n'(\zeta_1 a)$ is near zero. Let p denote a particular root of J_n or J_n'; then to zero order,

$$\begin{aligned} \zeta_1 a &= p \\ ha &= \beta_{nm} a = \beta_0 a(1 - \nu^2)^{1/2} \\ \zeta_2 a &= \beta_0 a(\epsilon' - i\epsilon'' - 1 - \nu^2)^{1/2} \end{aligned} \tag{A2}$$

where

$$\nu = p/\beta_0 a$$

Henceforth assume that

$$|\zeta_2 a| \gg |(4n^2 - 1)/8| \tag{A3a}$$

and

$$|\zeta_2 a| \gg |n| \tag{A3b}$$

It is convenient to postulate both inequalities, even though the first is more restrictive than the second unless $|n| = 1$ or $|n| = 2$.

If (A3a) is satisfied, the Hankel functions may be replaced by the first terms of their asymptotic expressions, and

$$\frac{H_n^{(2)\prime}(\zeta_2 a)}{H_n^{(2)}(\zeta_2 a)} = -i$$

Eq. (A1) becomes

$$\left(\zeta_1 a \tan\psi - \frac{nha}{\zeta_1 a}\right)^2 \frac{J_n(\zeta_1 a)}{J_n'(\zeta_1 a)} - (\beta_0 a)^2 \frac{J_n'(\zeta_1 a)}{J_n(\zeta_1 a)}$$
$$= \frac{i\zeta_1 a}{\zeta_2 a}\left[\left(\zeta_2 a \tan\psi - \frac{nha}{\zeta_2 a}\right)^2 + (\beta_0 a)^2(\epsilon' - i\epsilon'')\right]$$

It follows from (A3b), using the zero-order approximations (A2), that

$$| nha/\zeta_2 a | \ll | \beta_0 a(\epsilon' - i\epsilon'')^{1/2} |$$

so the characteristic equation finally takes the approximate form

$$\left(\zeta_1 a \tan \psi - \frac{nha}{\zeta_1 a}\right)^2 \frac{J_n(\zeta_1 a)}{J_n'(\zeta_1 a)} - (\beta_0 a)^2 \frac{J_n'(\zeta_1 a)}{J_n(\zeta_1 a)} = \frac{i\zeta_1 a}{\zeta_2 a}[(\zeta_2 a \tan \psi)^2 + (\beta_0 a)^2(\epsilon' - i\epsilon'')] \tag{A4}$$

Now let

$$\zeta_1 a = p + x, \qquad | x | \ll 1$$

where x is a small complex number. The normalized propagation constant becomes, to first order,

$$\begin{aligned} iha &= [(\zeta_1 a)^2 - (\beta_0 a)^2]^{1/2} \\ &= i\beta_0 a(1 - \nu^2)^{1/2} - i\nu x(1 - \nu^2)^{-1/2} \\ &= \alpha a + i(\beta_{nm} a + \Delta\beta a) \end{aligned}$$

where β_{nm} is the phase constant of the mode in a perfectly conducting guide, and the perturbation terms are

$$\alpha a + i\Delta\beta a = -\frac{i\nu x}{(1 - \nu^2)^{1/2}} \tag{A5}$$

For the TM_{nm} mode, let p be the m^{th} root of J_n ; then from Taylor's series, to first order in x,

$$J_n(\zeta_1 a) = J_n(p + x) = xJ_n'(p) \tag{A6}$$

Substituting (A6) into (A4), neglecting the first term on the left side of (A4), and replacing everything on the right side by its zero approximation according to (A2), one obtains

$$-\frac{(\beta_0 a)^2}{x} = \frac{ip\beta_0 a[(\epsilon' - i\epsilon'' - 1 + \nu^2)\tan^2 \psi + (\epsilon' - i\epsilon'')]}{(\epsilon' - i\epsilon'' - 1 + \nu^2)^{1/2}}$$

or

$$x = \frac{i(\xi + i\eta)}{\nu\left[1 + \left\{1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''}\right\}\tan^2 \psi\right]} \tag{A7}$$

where

$$\xi + i\eta = \frac{\left[1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''}\right]^{1/2}}{(\epsilon' - i\epsilon'')^{1/2}} \tag{A8}$$

It follows from (A5) and (A7) that for TM modes,

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}\left[1 + \left\{1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''}\right\}\tan^2\psi\right]} \tag{A9}$$

where $\xi + i\eta$ is given by (A8).

For the TE_{nm} mode, let p be the m^{th} root of J_n'; then

$$J_n'(\zeta_1 a) = J_n'(p + x) = \frac{(n^2 - p^2)x}{p^2} J_n(p)$$

Equation (A4) yields

$$x = \frac{ip^2\nu}{(p^2 - n^2)} \frac{\left[\tan\psi - \frac{n(1 - \nu^2)^{1/2}}{p\nu}\right]^2 (\xi + i\eta)}{\left[1 + \left\{1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''}\right\}\tan^2\psi\right]}$$

and, using (A5), we have for TE modes,

$$\alpha + i\Delta\beta$$

$$= \frac{p^2}{(p^2 - n^2)} \frac{\nu^2}{a(1 - \nu^2)^{1/2}} \frac{\left[\tan\psi - \frac{n(1 - \nu^2)^{1/2}}{p\nu}\right]^2 (\xi + i\eta)}{\left[1 + \left\{1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''}\right\}\tan^2\psi\right]} \tag{A10}$$

where $\xi + i\eta$ is given by (A8).

In view of (A5), the condition that $|x| \ll 1$ is equivalent to

$$\frac{(1 - \nu^2)^{1/2}}{\nu} |\alpha a + i\Delta\beta a| \ll 1 \tag{A11}$$

In all the numerical cases treated in the present paper, the approximate formulas agree well with the exact ones provided that the left side of (A11) is not greater than about 0.1.

A condition which is usually satisfied in practice, although not strictly a consequence of the assumptions (A3) or (A11), is

$$\left|\frac{1 - \nu^2}{\epsilon' - i\epsilon''}\right| \ll 1$$

This final approximation leads to the simple equations (7a) and (7b) of Section III, namely:

TM_{nm} *modes*

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}[1 + \tan^2\psi]}$$

TE_{nm} *modes*

$$\alpha + i\Delta\beta = \frac{(\xi + i\eta)}{a(1 - \nu^2)^{1/2}} \frac{\nu^2 p^2}{(p^2 - n^2)} \frac{[\tan\psi - n(1 - \nu^2)^{1/2}/p\nu]^2}{[1 + \tan^2\psi]}$$

where

$$\xi + i\eta = (\epsilon' - i\epsilon'')^{-1/2}$$

Wafer-Type Millimeter Wave Rectifiers*

By W. M. SHARPLESS

(Manuscript received June 18, 1956)

A wafer-type silicon point-contact rectifier and holder designed primarily for use as the first detector in millimeter wave receivers are described. Measurements made on a pilot production group of one hundred wafer rectifier units yielded the following average performance data at a wavelength of 5.4 millimeters: conversion loss, 7.2 db; noise ratio, 2.2; intermediate frequency output impedance 340 ohms. Methods of estimating the values of the circuit parameters of a point-contact rectifier are given in an Appendix.

INTRODUCTION

Point-contact rectifiers for millimeter waves have been in experimental use for several years. These units, for the most part, have been coaxial cartridges which were inserted in a fixed position, usually centered, in the waveguide. Impedance matching was accomplished by means of a series of matching screws preceding the rectifier and an adjustable waveguide piston following the rectifier. Tuning screws are generally undesirable because of the possibility of losses, narrow bandwidths and instability.

It is the purpose of this paper to describe a new type millimeter-wave rectifier and holder which were designed to eliminate the need for tuning screws and to provide a readily interchangeable rectifier of the flat wafer type. This wafer contains a short section of waveguide across which the point contact rectifier is mounted. The necessary low frequency output terminal (and the rectified current connection) together with the high-frequency bypass capacitor, are also contained within each wafer. The basic idea of the wafer-type rectifier is that the unit can be inserted in its holder and moved transversely to the waveguide to obtain a resistive match to the guide; the reactive component of the rectifier impedance is then tuned out by an adjustable waveguide plunger behind the rectifier.

* This work was supported in part by Contract Nonr-687(00) with the Office of Naval Research, Department of the Navy.

The wafer unit and holder were developed primarily for use as the first converter in double detection receivers operating in the 4- to 7-millimeter wavelength range. In order to check the practicability of the design and to supply rectifiers for laboratory use, a pilot production group of one hundred units was processed and measured. Performance data obtained with this group are presented. A balanced converter using wafer rectifiers is also described.

Methods of estimating the values of the various circuit parameters of a point-contact rectifier are outlined in an appendix. These calculations proved useful in the design of the wafer unit and in predicting the broadband performance of the converter.

DESCRIPTION OF WAFER UNIT AND HOLDER

Fig. 1 is a drawing of the wafer type rectifier. The unit is made from stock steel $\frac{1}{16}$-inch thick and is gold plated after the milling, drilling and soldering operations are completed. To allow for the transverse impedance matching adjustment, the section of waveguide contained in the wafer is made wider than the RG98U input guide to the holder. By making the wafer thin ($\frac{1}{16}$ inch), the short sections of unused guide on either side will remain "cut-off" over the operating range of the rectifiers. The silicon end of the rectifier consists of a copper pin on which the silicon is press mounted, the assembly held in place with Araldite cement which also serves as the insulating material for a quarter-wavelength long high frequency bypass capacitor. The pin serving as the intermediate frequency and direct current output lead is also cemented in place with Araldite cement. A soft solder connection is made between this pin and the pin holding the silicon wafer. A nickel pin with a conical end on which a pointed tungsten contact spring is welded is pressed into place from the opposite side of the guide at the time of final assembly.

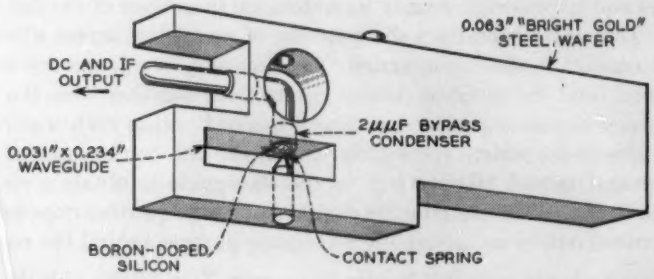


Fig. 1 — Millimeter-wave wafer unit.

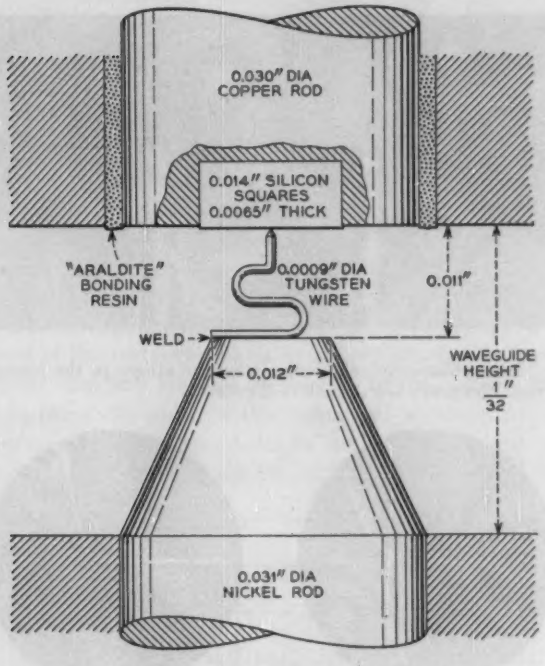


Fig. 2 — Millimeter-wave point-contact assembly.

The region of the wafer unit containing the silicon and point contact is shown in Fig. 2. The methods used in preparing the silicon wafer and the spring contact point are similar in many respects to the standard techniques used in the manufacture of rectifiers for longer wavelengths. Some modifications and refinements in technique are called for by a decrease in size and the increased frequency of operation.

A single-crystal ingot, grown from high purity DuPont silicon doped with 0.02 per cent boron, furnishes the material for the silicon squares used in the wafer unit. Slices cut from the ingot are polished and heat treated. Gold is evaporated on the back surface and the slices are diced into squares approximately 0.014-inch square and 0.0065-inch thick. These squares are pressed into indentations formed in the ends of the 0.030-inch copper pins which have previously been tin-plated. The rods are then cemented in place in the wafer. The spring contact points are made of pure tungsten wire that has been sized to 0.9 mil in diameter by an electrolytic etching process. A short length of this wire is spot welded on the conical end of the 0.031-inch nickel rod. The wire is then bent into

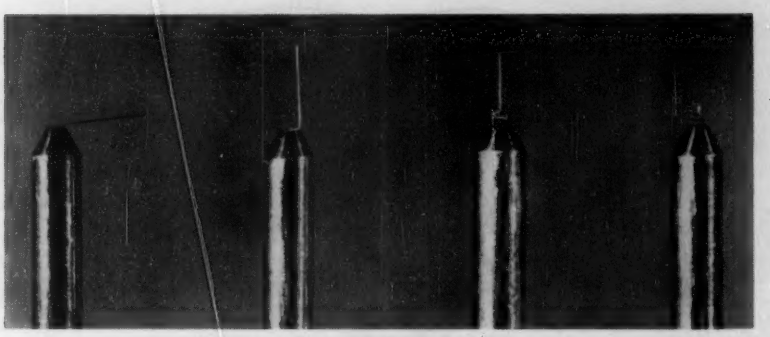


Fig. 3 — Micro-photograph showing successive stages in the formation of the contact spring. The posts are $\frac{1}{16}$ inch in diameter.

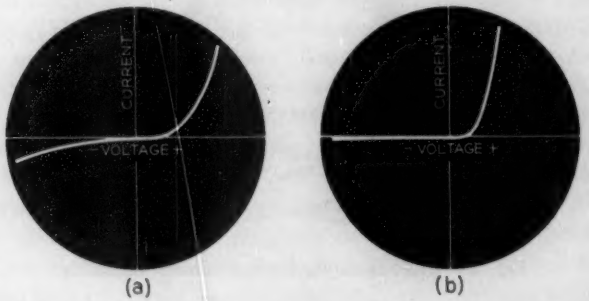


Fig. 4 — Cathode-ray oscilloscope display of wafer unit static characteristic: (a) before and (b) after tapping.

the "S" configuration in a forming jig. By an electrolytic process the spring is then cut to the proper length and pointed. The micro-photographs in Fig. 3 show successive stages in the formation of the contact spring.

In the final assembly of the unit the nickel rod with the contact spring is pressed into place until contact is made with the silicon. It is then advanced a half mil to obtain the proper contact pressure. The voltage-current characteristics as viewed at 60 cycles on a cathode-ray oscilloscope will then appear as shown in Fig. 4(a). The unit is "tapped" into final adjustment. This is done by clamping the unit in a holder and rapping it sharply on the top of a hard wood bench. This procedure requires experience as excessive "tapping" will impair the performance of the unit. Usually one vigorous "tap" is sufficient to produce the desired effect and the voltage-current characteristic will appear as

shown in Fig. 4(b). The static characteristic of a typical unit is shown in Fig. 5.

The conversion loss of each unit is measured before the end of the nickel rod carrying the contact point is cut off flush with the wafer. In the event that this initial measurement shows that the conversion loss exceeds an arbitrarily chosen upper limit (8.5 db), it is possible at this stage to withdraw the point and replace it with a new one. This procedure, which was necessary on only a few of the units processed, always resulted in an acceptable unit. The final operation is to cut off the protruding end of the nickel rod flush with the wafer.

A holder designed to use the wafer units is shown in Figs. 6 and 7. At the input end of the converter block is a short waveguide taper section to match from standard RG98U waveguide to the $\frac{1}{32}$-inch high waveguide used in the wafer unit. As the wafer unit is moved in and out to match the conductance of the crystal to the waveguide, the output pin of the wafer unit slides in a chuck on the inner conductor of the coaxial

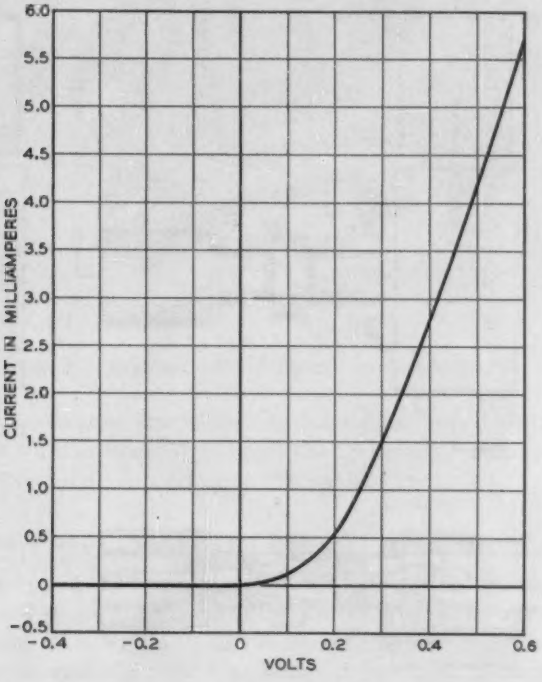


Fig. 5 — Static characteristic of typical millimeter-wave wafer unit.

output jack. The unit may be clamped in position after matching adjustments are made by tightening the knurled thumb screw which pushes a cylindrical slug containing an adjustable piston against the wafer unit. The piston is a short septum which slides in shallow grooves in the top and bottom of the $\frac{1}{32}$-inch high waveguide, thus dividing the waveguide into two guides which are beyond cut-off. This septum is made of two pieces of thin beryllium copper bowed in opposite directions so that good contact is made to the sides of the grooves in the top and bottom of the waveguide. Since the piston with its connecting rod is very light in weight and is held firmly in place by the spring action of the bowed septum, no additional locking mechanism need be provided. Since the rectifier is essentially broadband by design, the adjustment of the piston is not critical and is readily made by hand. The piston rod is protected by a cap which is snapped in place over the thumb screw when all tuning adjustments are completed.

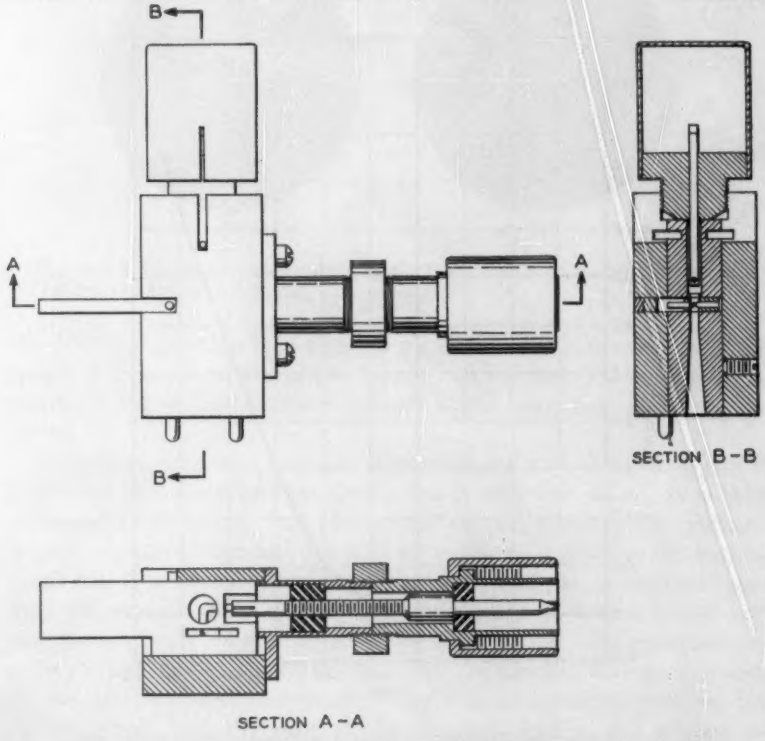


Fig. 6 — Assembly drawing of millimeter-wave converter.

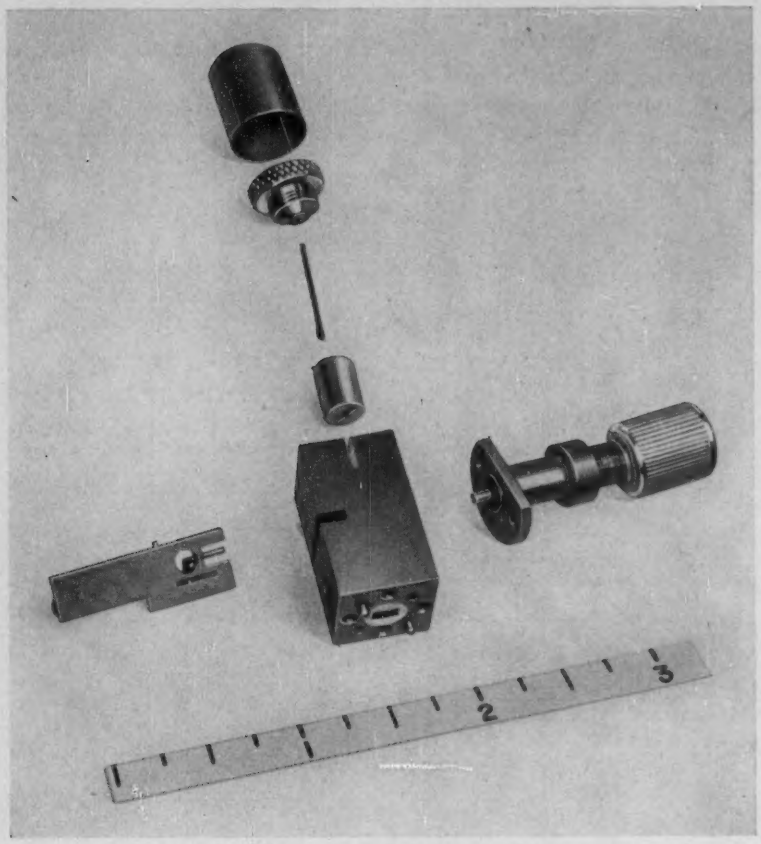


Fig. 7 — Exploded view of millimeter-wave converter.

With the converter fixed-tuned at 5.4 millimeters, a shift in wavelength to 6.3 millimeters (17 per cent change) produces a mismatch loss of from 1.6 to 4.0 db depending on the rectifier used.

PERFORMANCE DATA FOR WAFER-TYPE RECTIFIER UNIT

A pilot group of one hundred wafer units was processed and measured. Figs. 8, 9 and 10 are bar graphs of the distribution of the conversion loss L, and noise ratio N_R*, and the 60 megacycle intermediate frequency output impedance Z_{IF}, for the hundred rectifiers measured in the

* N_R is the ratio of the noise power available from the rectifier to the noise power available from an equivalent resistor at room temperature.

mixer of Fig. 7 at a wavelength of 5.4 millimeters. In order that the measurements might be more readily compared with those made on commercially available rectifiers used at longer wavelengths, the available beating oscillator power was maintained at a level of one milliwatt for all measurements.* Further, in the case of the conversion loss, a

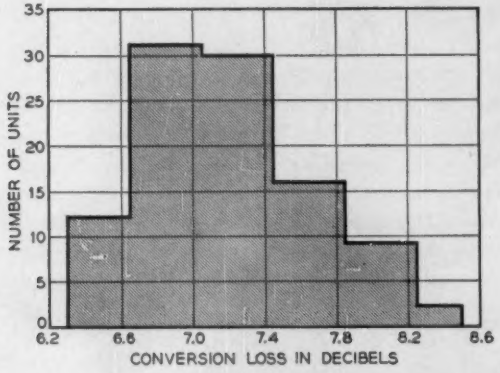


Fig. 8 — Conversion loss (L) of 100 wafer units at a wavelength of 5.4 millimeters with one-milliwatt beating oscillator drive (average 7.2 db).

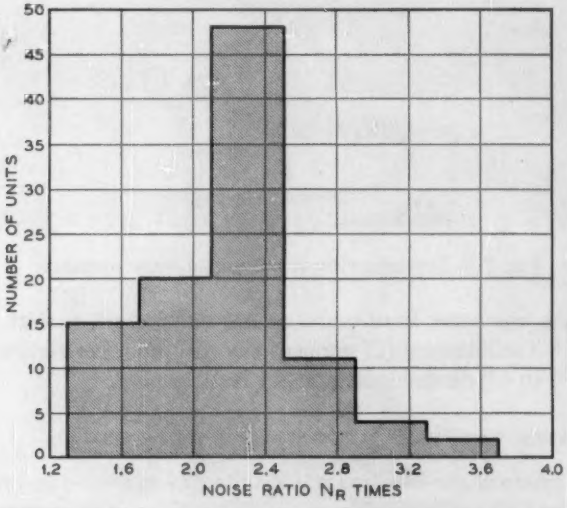


Fig. 9 — Noise Ratio (N_R) for 100 wafer units at a wavelength of 5.4 millimeters with a one-milliwatt beating oscillator drive (average 2.21 times).

* Power levels were determined by the use of a calorimeter. See, A Calorimeter for Power Measurements at Millimeter Wavelengths, I. R. E. Trans., MTT-2, pp. 45–47, Sept., 1954.

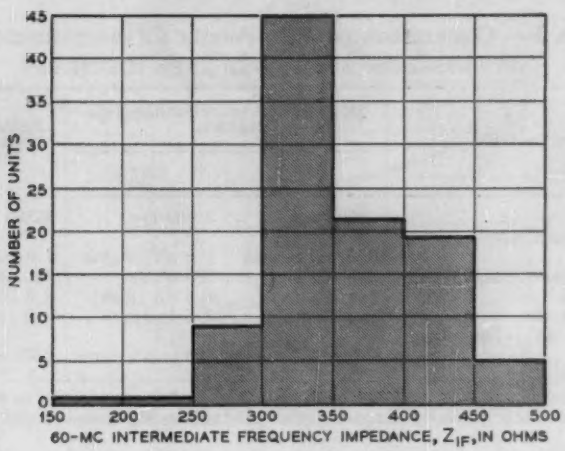


Fig. 10 — Sixty-megacycle intermediate-frequency output impedance (Z_{IF}) for 100 wafer units with one milliwatt beating oscillator drive (average 338 ohms)

limit of 8.5 db was arbitrarily adopted. This required the readjustment of eleven units, with a new point inserted in each case. No units were rejected because of high noise and none of the hundred units processed was lost.

From the bar graphs it may be seen that the wafer units have the average characteristics shown in the accompanying table at a wavelength of 5.4 millimeters.*

Conversion Loss L	7.2 db (5.3 times)
Noise Ratio N_R	2.2 times
IF Impedance (60 mc) Z_{IF}	338 ohms

Knowing the noise figure, N_{IF}, of the *IF* amplifier intended for use with the rectifiers, the overall receiver noise figure, N_{REC}, may be calculated by the following formula (using numerical ratios):

$$N_{REC} = L(N_R - 1 + N_{IF})$$

Assuming an *IF* amplifier noise figure of 4.0 db ($2\frac{1}{2}$ times) and the average values of "L" and "N_R" given above for the millimeter wafer units, we have for the case of a noiseless beating oscillator;

$$N_{REC} = 5.3\ (2.2 - 1 + 2.5) \approx 20\ (13\ \text{db})$$

* A few wafer units have also been measured at a wavelength of 4.16 millimeters. The conversion losses averaged about 1.6 db greater than those measured at a wavelength of 5.4 millimeters.

TABLE I — COMPARISON OF LOW-POWER CHARACTERISTICS OF CARTRIDGE-TYPE AND WAFER-TYPE RECTIFIERS

Test Conditions	JAN Specifications for Cartridge-Type Rectifiers		Performance of Wafer-Type Rectifiers
	IN26	IN53	
Frequency	23984 mc	34860 mc	55500 mc
Beating oscillator power level	1.0 milliwatts	1.0 milliwatts	1.0 milliwatts
Noise reference resistor	300 ohms	300 ohms	300 ohms
Conversion loss	8.5 db (max)	8.5 db (max)	8.5 db (max)*
Noise ratio	2.5 (max)	2.5 (max)	2.2 (average)†
Nominal IF impedance range	300 to 600 ohms	400 to 800 ohms	250 to 500 ohms

* Limit arbitrarily set on basis of 100 per cent yield as explained in the text.
† Limit not set. Actually in more recent production N_R has averaged 1.7 times.

In practice, the beating oscillator noise sidebands can be eliminated by the use of a matched pair of rectifiers in a balanced converter arrangement described later. The resulting overall noise figure of 13 db on an average compares quite favorably with the figures obtained at longer wavelengths.

In Table I it is seen that a high percentage of the group of one hundred units would be able to pass low-power JAN specifications similar to those set down for the commercially available IN26 and IN53 rectifiers used at longer wavelengths.

EFFECT OF VARYING THE BEATING OSCILLATOR POWER

When the optimum over-all receiver noise figure is desired, it may well turn out that a beating oscillator drive of one milliwatt (corresponding to a dc rectified current for different wafers of from $\frac{9}{10}$ to $1\frac{1}{4}$ milliamperes) is too large. Fig. 11 shows the effect on the performance of a typical unit as the beating oscillator drive is varied above and below the one milliwatt level as indicated by the change in the dc rectified current. It is seen that the value of N_R tends to increase rapidly for a beating oscillator drive much in excess of one milliwatt; with reduced drive, the over-all noise figure of the receiver, N_{REC} for the example taken, improves, reaching a minimum value near a rectified current of about $\frac{7}{10}$ milliampere corresponding to a drive of about $\frac{2}{3}$ of a milliwatt.

A BALANCED CONVERTER FOR WAFER UNITS

A broad-band balanced first converter has been developed which makes use of a pair of wafer-type millimeter-wave rectifiers. This converter

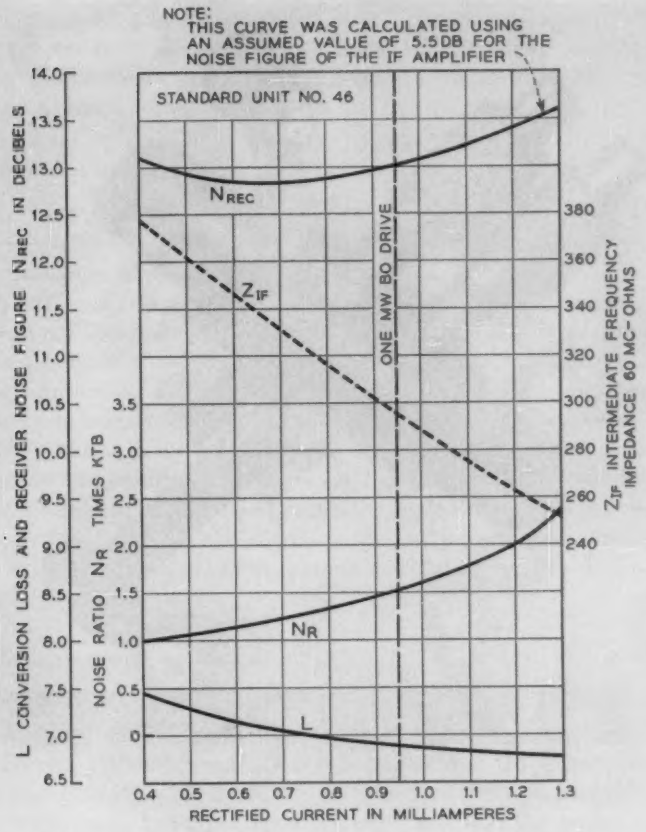


Fig. 11 — Typical performance curves for wafer-type rectifiers.

was designed to operate over the 4- to 7-millimeter band and is pictured in Fig. 12. A compact arrangement has been achieved which makes use of a waveguide finline-to-coaxial input circuit for the beating oscillator while the signal is introduced through a separate impedance-matched waveguide "Tee" section. Return loss measurements show that with a matched pair of wafer units, fixed-tuned in the center of the 5- to 6-millimeter band, an excess loss of about 1 db may be expected at the edges of a 15 per cent band. At midband, an improvement of 5 db in over-all receiver noise figure was obtained by substituting the balanced converter for an unbalanced one in a test receiver using an M1805 milli-

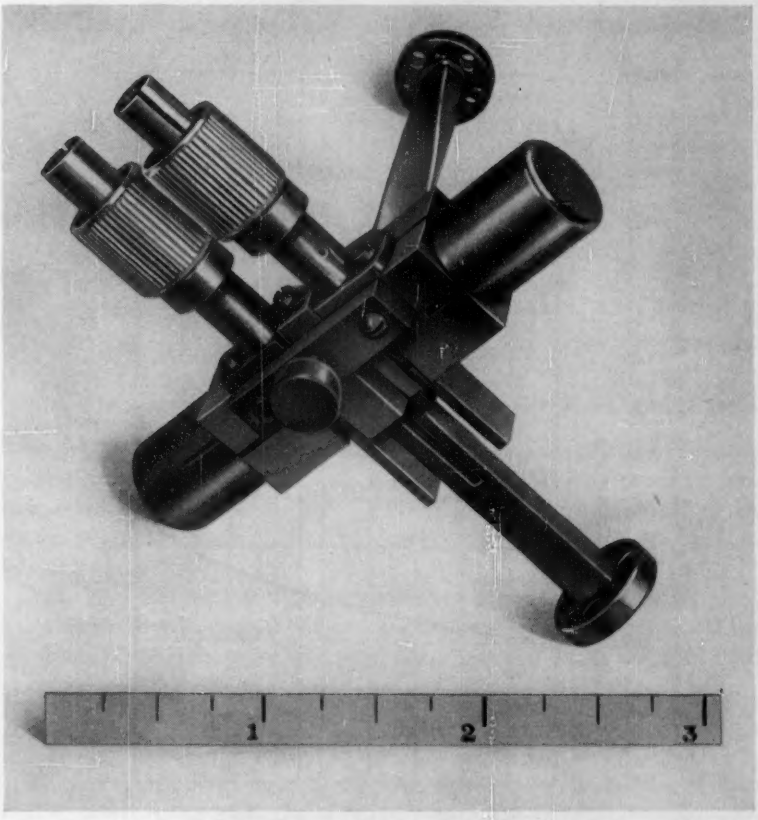


Fig. 12 — Balanced converter with wafer-type rectifiers.

meter-wave reflex klystron* as the beating oscillator and a 60-mc intermediate frequency amplifier with a 5-db noise figure.

REVERSED POLARITY WAFER UNIT

When using crystal rectifiers in a balanced converter arrangement, there is a distinct advantage, circuit-wise, in using two units of opposite polarity. For this reason, a reversed-polarity wafer type rectifier has also been developed. This was done by interchanging the silicon and the

* E. D. Reed, A Tunable, Low-Voltage Reflex Klystron for Operation in the 50 to 60 kmc Band, B. S. T. J., **34**, pp. 563–599, May, 1955.

point contact spring in a standard unit. The standard and reverse-polarity wafer have the same outer physical dimensions and thus they may be used interchangeably in the holders as dictated by the specific problems at hand.

CONCLUDING REMARKS

Aside from their intended use as first detectors in double detection receivers, wafer units have been used for single detection measurements at frequencies as high as 107 kmc.

It is felt that the pilot production group of one hundred units is a sample of sufficient size to yield representative data and to demonstrate the practicability of the design. It should be pointed out that the units have not been filled with protective waxes and have not been subjected to temperature-humidity cycling tests. However, a few reference units have been in use in the laboratory for over a year and have shown no measurable deterioration. No attempt has been made to establish a burn-out rating for the rectifier, but units have withstood available cw input powers of the order of 15 milliwatts and narrow pulse discharges of the order of $\frac{1}{10}$ erg without causing noticeable changes in the conversion loss or noise ratio.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to H. T. Friis and A. B. Crawford for their helpful suggestions and guidance during the course of this work. Extensive use has also been made of the experience and techniques of R. S. Ohl. E. F. Elbert participated in the development of the wafer unit, being particularly concerned with the techniques of fabrication. H. W. Anderson and S. E. Reed were most helpful in solving mechanical problems encountered in the production of wafer units and holders.

APPENDIX

This section describes some calculations that were made for the purpose of estimating the values of the various parameters involved in the design of a high frequency point contact rectifier. These parameters are the barrier resistance, the spreading resistance, the capacitance of the barrier layer and the inductance of the contact spring. Knowing the approximate values of these parameters one can, by an equivalent circuit analysis, arrive at a simple parallel circuit for the rectifier which may

be used in designing an appropriate holder. Also, using this equivalent circuit, one may calculate the bandwidth expected for the converter.

Fig. 13 shows the point contact rectifiers under consideration and an enlarged view of the point contact region. On the right of the figure are shown equivalent circuits of the rectifier. Circuit I is the generally accepted circuit of a point contact rectifier. The true circuit for a rectifier operating at millimeter wavelengths is probably more complicated than that shown in the figure but, for an approximate analysis, the simplified circuit has been found to yield useful results. In the following paragraphs, values are derived for the parameters of this equivalent circuit. MKS units are used and values appropriate to the millimeter wave wafer unit are used as examples.

Spreading Resistance

The spreading resistance, R_s, may be calculated if we know the resistivity of the silicon used for the rectifier and the radius of the contact area formed when the units are assembled. For DuPont high-purity silicon, doped with 0.02 per cent boron by weight, W. Shockley* gives the resistivity, ρ, as 0.90×10^{-4} ohm meters. From numerous measurements on millimeter wave contact areas, R. S. Ohl finds the contact radius, r_1, to be about 1.25×10^{-6} meters. The spreading resistance,

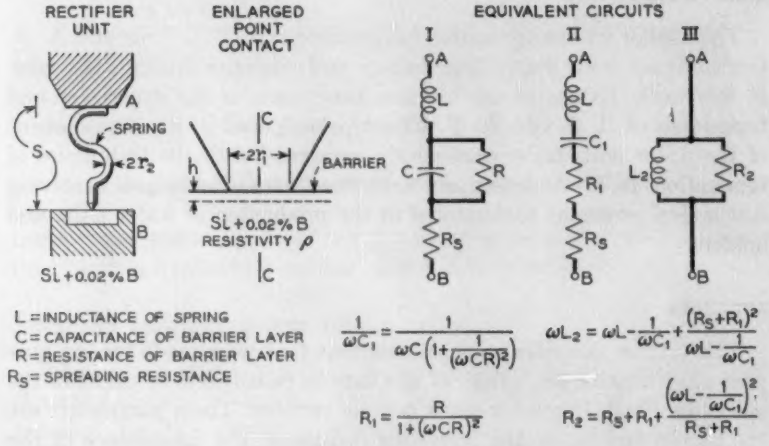


Fig. 13 — Point contact rectifier and equivalent circuits.

* W. Shockley, Electrons and Holes in Semiconductors, New York: D. Van Nostrand Co., Inc., 1950, p. 284.

R_s, assuming a circular contact area, may be calculated from the formula, $R_s = \rho/4r_1$.* For the above example, $R_s = 18$ ohms.

Barrier Resistance

The approximate operating value of the barrier resistance, R, may be determined from a knowledge of the intermediate frequency impedance of a typical rectifier. A. B. Crawford has shown that the optimum intermediate frequency output impedance of a crystal mixer rectifier is a function of the exponent of the static characteristic of the rectifier and the impedance presented to the rectifier at the image and signal frequencies. This information is presented in Fig. 12.3–6 in G. C. Southworth's book.† In the millimeter wave case it is a good assumption that the impedances for the signal and image frequencies are equal; for this case and for matched conditions, the magnitude of the high frequency impedance is seen to be a simple multiple of the intermediate frequency impedance R_{IF}.

From numerous measurements on mixer rectifiers operating at different frequencies it is known that the intermediate frequency impedance of an average rectifier is very nearly 400 ohms. We also know from the DC static characteristics of our millimeter wave type rectifiers that the average exponent is about four. With this information, and the curves in Southworth's book, it is found that $R \approx R_{IF}/1.5$. Thus, the barrier resistance R is about 250 ohms.‡

Capacitance of Barrier Layer

From a knowledge of the point contact area, the barrier layer thickness, and the dielectric constant of the silicon, the capacitance of the point contact may be calculated. The radius of the contact point area is the same as that used for the calculation of the spreading resistance. The barrier layer thickness, h, for the heat treated silicon used for millimeter waves has been measured by R. S. Ohl to be about 10^{-8} meters. The dielectric constant of silicon is $\varepsilon_r = 13$. The capacitance is given by the following formula

$$C = \frac{r_1^2 \varepsilon_r}{3.6h \times 10^{10}} \text{ farads} \qquad (1)$$

* J. H. Jeans, Mathematical Theory of Electricity and Magnetism, 5th Ed., Cambridge University Press, 1925.

† G. C. Southworth, Principles and Applications of Waveguide Transmission, New York: D. Van Nostrand Co., Inc., 1950.

‡ This resistance cannot be readily measured directly at millimeter waves.

For the above case $C = 5.7 \times 10^{-14}$ farads or $1/\omega C$ at 5.4 millimeters is about 50 ohms.

The accuracy of this capacitance calculation can be verified later when a completed rectifier is measured for its high frequency conversion loss. This is possible because we know the calculated low frequency conversion loss of the rectifier, for the case of zero spreading resistance from Southworth's book, Fig. 12.3–7. For an exponent of four this loss is given as 4.4 db. The additional loss at high frequency due to the capacitance, C, may be calculated (see Equivalent Circuit II) by the formula:

$$\text{Additional Loss} = 10 \log_{10} \frac{R_1 + R_s}{R_1} \text{ db} \tag{2}$$

From the text (Fig. 8), it is seen that the average wafer rectifier unit has a conversion loss at 5.4 millimeters of 7.2 db; thus, the difference between the low and high frequency conversion losses is very nearly 3 db. This means that about one-half the signal power is lost in the spreading resistance; hence R_1 and R_s are about equal. By transferring back to Equivalent Circuit I, the average value of the capacitance is found to be 4.1×10^{-14} farads, which is a reasonable check with the calculated value given by (1).

Inductance of the Contact Spring

The remaining parameter of the equivalent circuit to be determined is the inductance of the contact spring. The value of the equivalent parallel resistance, R_2, depends on the inductance L (the other parameters being fixed), or conversely, for a given value of R_2, the appropriate value for L may be calculated from the formula for Equivalent Circuit III. For an off-center match of the rectifier to the waveguide, R_2 must equal the guide impedance, Z_d, at the rectifier location. Also, for a match, the distance, ℓ, from the rectifier to the waveguide piston must

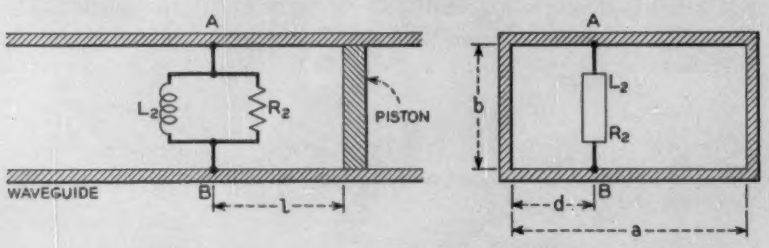


Fig. 14 — Matching circuit for rectifier offset in waveguide.

satisfy the relation, $Z_d \tan 2\pi l/\lambda_g = -\omega L_2$. (See Fig. 14.) The impedance of the guide as a function of d/a is given by,

$$Z_d = 240\pi \frac{b}{a} \frac{1}{\sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}} \sin^2 \frac{\pi d}{a} \tag{3}$$

As a compromise between electrical and mechanical requirements, a waveguide height, b, of $\frac{1}{32}$ inch was chosen for the wafer unit; the width of the guide was taken to be the same as *RG98U*. For $b = 7.88 \times 10^{-4}$, $a = 3.76 \times 10^{-3}$, $d/a = \frac{1}{4}$ and $\lambda = 5.4 \times 10^{-3}$, (3) gives a value of 113 ohms for Z_d (and R_2). The appropriate value for L then becomes 3.38×10^{-10} henries.

An estimate of the size of a contact spring having the inductance given above can be made from the formula below which gives the inductance of a straight thin wire of length S as a function of its sidewise position in the waveguide.* (See Fig. 15.)

$$L = 2S \log_e \frac{2a \sin \frac{\pi d}{a}}{\pi r_2} \times 10^{-7} \text{ henries} \tag{4}$$

$$r_2 \ll d$$

For $d/a = \frac{1}{4}$ and $2r_2 = 2.28 \times 10^{-5}$ (0.9×10^{-3} inches), the length, S, is found to be about 3.38×10^{-4} meters or about 0.013 inches.

Since the spring must be so very small, the circuit from the base of the spring to the waveguide wall is completed with a large low inductance conical post as shown in Fig. 2 of the text.

Bandwidth Calculation

Having assigned values to all the parameters of the equivalent circuit, it is now possible to calculate the mismatch loss of a fixed-tune

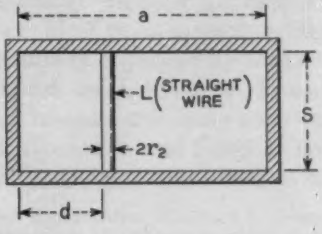


Fig. 15 — Thin wire in waveguide.

* Private communication from S. A. Schelkunoff.

converter for a given change in operating wavelength. This loss is given by the following formula:

Mismatch loss

$$= 10 \log_{10} \frac{4Z_d}{R_2\left[\left(1+\frac{Z_d}{R_2}\right)^2+\left(\frac{Z_d}{\omega L_2}+\frac{1}{\tan 2\pi\ell/\lambda_g}\right)^2\right]} db \quad (5)$$

For the wafer unit, calculation shows that the rectifier is matched to the waveguide at a wavelength of 5.4×10^{-3} meters for $d/a = \frac{1}{4}$ and $\ell = 3.14 \times 10^{-3}$. If now the wavelength is changed to 6.3×10^{-3} meters, without retuning (17 per cent change) the mismatch loss calculated by (5) is 1.6 db. It was stated in the text that a number of wafer units gave measured mismatch losses of from 1.6 to 4.0 db for a 17 per cent change in wavelength without retuning. This is considered to be a reasonable correlation between calculations and measurements.

Frequency Conversion by Means of a Nonlinear Admittance

C. F. EDWARDS

(Manuscript received June 20, 1956)

This paper gives a mathematical analysis of a heterodyne conversion transducer in which the nonlinear element is made up of a nonlinear resistor and a nonlinear capacitor in parallel. Curves are given showing the change in admittance and gain as the characteristics of the nonlinear elements are varied. The case where a conjugate match exists at the terminals is treated.

It is shown that when the output frequency is greater than the input frequency, modulators having substantial gain and bandwidth are possible, but when the output frequency is less than the input frequency, the converter loss is greater than unity and is little affected by the nonlinear capacitor. The conditions under which a conjugate match is possible are specified and it is concluded that a nonlinear capacitor alone is the preferred element for modulators and that a nonlinear resistor alone gives the best performance in converters.

INTRODUCTION

Point contact rectifiers using either silicon or germanium are used as the nonlinear element in microwave modulators to change an intermediate frequency signal to an outgoing microwave signal and in receiving converters to change an incoming microwave signal to a lower intermediate frequency. Most point contact rectifiers now in use behave as pure nonlinear resistors as evidenced by the fact that in either of the above uses the conversion loss is the same. In recent experiments with heterodyne conversion transducers* using point contact rectifiers made with ion bombarded silicon this was found to be no longer true. The conversion loss of the modulator was found to be unusually low and

* This term is defined in American Standard Definitions of Electrical Terms — ASA C42 — as "a conversion transducer in which the useful output frequency is the sum or difference of the input frequency and an integral multiple of the frequency of another wave".

that of the converter was several decibels greater. In one instance the loss in a modulator used to convert a 70 mc signal to one at 11,130 mc was found to be only 2.3 db but when the direction of transmission through it was reversed and it was used as a converter, the loss was 7.8 decibels.

Similar effects were observed several years ago in conversion transducers using welded contact germanium rectifiers.[1] In these early experiments substantial converter gain and negative conductance at the intermediate frequency terminals were also observed. These results were accounted for by assuming the presence of a nonlinear capacitance at the point contact in parallel with the nonlinear resistance. At that time attention was devoted mainly to the behavior of converters where noise is a vital factor. It was found that although the conversion loss could be reduced, the noise temperature increased and no improvement in noise figure resulted. However, the noise temperature requirements in modulators are much less severe and the nonlinear capacitance effect is useful and can substantially improve the performance.

THEORY

The mathematical analysis given here was undertaken in order to clarify the effect of the nonlinear capacitance in the frequency conversion process and to obtain an estimate of the usefulness of modulators exhibiting gain. The analysis is restricted to the simplest case in which signal voltages are allowed to develop across the nonlinear elements at the input and output frequencies only. This is not an unrealistic restriction since the conversion transducers used in microwave relay systems have filters associated with them which suppress the modulation products outside the signal band. The final results will be given only for those conditions which permit a conjugate match at the input and output of the transducer.

The procedure used to obtain expressions for the admittance and gain of conversion transducers utilizing a nonlinear element made up of a nonlinear resistance and a nonlinear capacitance in parallel follows the commonly used method of treating the nonlinear elements as local oscillator controlled linear time varying elements.[2] The current through the nonlinear resistor is a function of the applied voltage. The derivative of this function is the conductance as a function of the applied voltage. Thus when the local oscillator is applied, the conductance varies at the local oscillator frequency and the conductance as a function of time may be obtained. This is periodic and may be expressed as a Fourier series. The conductance is real and if we make the usual assumption that

it may be expressed as an even function of time, we may write

$$\gamma = \cdots + G_2 e^{-j2\omega_0 t} + G_1 e^{-j\omega_0 t} + G_0 + G_1 e^{j\omega_0 t} + G_2 e^{j2\omega_0 t} + \cdots \quad (1)$$

where $\omega_0/2\pi$ is the local oscillator frequency f_0 and the Fourier coefficients G_n are real. Similarly the charge on the nonlinear capacitor is a function of the applied voltage. The derivative of this function is the capacitance as a function of the applied voltage. The application of the local oscillator thus causes the capacitance to vary at the local oscillator frequency so that it also may be expressed as a Fourier series. The capacitance κ is real, and assuming it may be expressed as an even function of time, we have

$$\kappa = \cdots + C_2 e^{-j2\omega_0 t} + C_1 e^{-j\omega_0 t} + C_0 + C_1 e^{j\omega_0 t} + C_2 e^{j2\omega_0 t} + \cdots \quad (2)$$

It is assumed that the current and charge functions are single valued and that their derivatives are always positive.

When a small signal voltage v is applied to the nonlinear resistor, the signal current through the resistor is given by γv. When it is applied to the nonlinear capacitor the charge on the capacitor is κv. The total current i which flows through the two nonlinear elements connected in parallel thus becomes

$$i = \gamma v + \frac{d}{dt}(\kappa v) \quad (3)$$

v of course must be small and not affect the value of γ and κ.

Fig. 1 shows a heterodyne conversion transducer made up of a nonlinear resistor and a nonlinear capacitor in parallel driven by an internal local oscillator. f_1 is the signal frequency at the terminals 1–2, and y_1 is the external admittance connected to these terminals. The signal frequency at the terminals 3–4 is f_2, and y_2 is the external admittance.

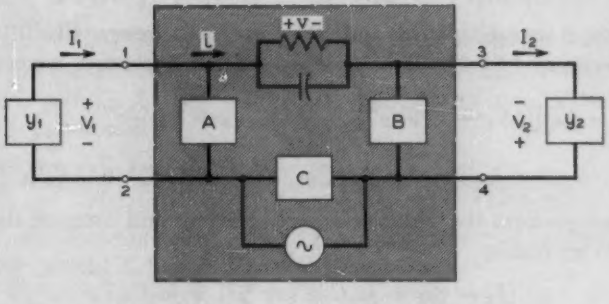


Fig. 1 — Heterodyne conversion transducer.

A, B and C are ideal frequency selective networks whose admittances are zero at f_1, f_2 and f_0 respectively, and infinite at all other frequencies. This circuit permits the application of the local oscillator voltage to the nonlinear elements but permits signal voltages to develop across them at f_1 and f_2 only. Similarly, signal currents at frequencies other than f_1 and f_2 encounter no external impedance, so they cannot alter the signal voltage or contribute to the external power. This, of course, assumes that if the nonlinear element is a point contact rectifier the spreading resistance normally present is negligible.

If f_1 is a frequency less than half the local oscillator frequency f_0 (it is generally very much less), the network B can be selected to make f_2 either $f_0 + f_1$, or $f_0 - f_1$. To distinguish between the two cases, we will call the former a noninverting conversion transducer since an increase in one signal frequency causes an increase in the other. The latter will be called an inverting conversion transducer since an increase in one signal frequency results in a decrease in the other. When y_1 contains the generator and y_2 the load, the device becomes a modulator. When y_2 contains the generator and y_1 the load, it is a converter.

The real part of the signal voltage may be written

$$v = V_1 e^{j\omega_1 t} + V_1^* e^{-j\omega_1 t} + V_2 e^{j\omega_2 t} + V_2^* e^{-j\omega_2 t} \tag{4}$$

where V^* is the complex conjugate of V and $\omega = 2\pi f$. Similarly, the real part of the signal current may be written

$$i = I_1 e^{j\omega_1 t} + I_1^* e^{-j\omega_1 t} + I_2 e^{j\omega_2 t} + I_2^* e^{-j\omega_2 t} \tag{5}$$

If we multiply equations (1) and (4) and retain only those terms containing f_1 and f_2 we obtain, in the case of the non-inverting conversion transducer where $f_2 = f_0 + f_1$,

$$\begin{aligned} \gamma v = {} & [G_0 V_1 + G_1 V_2] e^{j\omega_1 t} + [G_1 V_1 + G_0 V_2] e^{j\omega_2 t} \\ & + [G_0 V_1^* + G_1 V_2^*] e^{-j\omega_1 t} + [G_1 V_1^* + G_0 V_2^*] e^{-j\omega_2 t} \end{aligned} \tag{6}$$

Similarly, if we multiply (2) and (4) we get an expression like (6) with the G's replaced by C's. If we differentiate this expression we get

$$\begin{aligned} \frac{d}{dt}(\kappa v) = {} & j\omega_1 [C_0 V_1 + C_1 V_2] e^{j\omega_1 t} + j\omega_2 [C_1 V_1 + C_0 V_2] e^{j\omega_2 t} \\ & - j\omega_1 [C_0 V_1^* + C_1 V_2^*] e^{-j\omega_1 t} - j\omega_2 [C_1 V_1^* + C_0 V_2^*] e^{-j\omega_2 t} \end{aligned} \tag{7}$$

When we perform the addition indicated by (3) and compare the result with (5) we obtain

$$\begin{aligned} I_1 &= [G_0 + j\omega_1 C_0] V_1 + [G_1 + j\omega_1 C_1] V_2 \\ I_2 &= [G_1 + j\omega_2 C_1] V_1 + [G_0 + j\omega_2 C_0] V_2 \end{aligned} \tag{8}$$

Going through the same steps for the inverting conversion transducer where $f_2 = f_0 - f_1$ we obtain

$$\begin{aligned} I_1 &= [G_0 + j\omega_1 C_0]V_1 + [G_1 + j\omega_1 C_1]V_2^* \\ I_2^* &= [G_1 - jw_2 C_1]V_1 + [G_0 - j\omega_2 C_0]V_2^* \end{aligned} \tag{9}$$

Equations (8) and (9) are in the form

$$\begin{aligned} I_1 &= Y_{11}V_1 + Y_{12}V_2 \\ I_2 &= Y_{21}V_1 + Y_{22}V_2 \end{aligned} \tag{10}$$

A heterodyne conversion transducer may thus be represented by a linear 4-pole, and the admittance and gain of the 4-pole may be expressed in terms of the admittance coefficients. In Fig. 1 we see that the admittance of the 4-pole y_1' at the terminals 1–2 is equal to I_1/V_1 and the admittance y_2 connected to terminals 3–4 is $-I_2/V_2$. Putting these in (10) we find

$$y_1' = Y_{11} - \frac{Y_{12}Y_{21}}{Y_{22} + y_2} \tag{11}$$

Similarly the admittance of the 4-pole y_2' at the terminals 3–4 is I_2/V_2 and the admittance y_1 connected to terminals 1–2 is $-I_1/V_1$. Putting these in (10) gives

$$y_2' = Y_{22} - \frac{Y_{12}Y_{21}}{Y_{11} + y_1} \tag{12}$$

To compute the gain of the 4-pole when y_1 contains the generator and y_2 the load, it is convenient to assume a current source connected across y_1. If the current from this source is I_0 we have $I_1 = I_0 - y_1V_1$. I_2 equals $-y_2V_2$ as before. Putting these in (10) gives

$$\frac{I_0}{V_2} = Y_{12} - \frac{(Y_{11} + y_1)(Y_{22} + y_2)}{Y_{21}} \tag{13}$$

If we let $y_1 = g_1 + jb_1$ and $y_2 = g_2 + jb_2$, the power in the load is $V_2^2 g_2$ and the power available from the generator is $I_0^2/4g_1$. Therefore the transducer gain Γ_{12} defined as the ratio of the power in y_2 to that available from y_1 becomes

$$\Gamma_{12} = 4g_1g_2\frac{V_2^2}{I_0^2} = 4g_1g_2\left|\frac{Y_{21}}{Y_{12}Y_{21} - (Y_{11} + y_1)(Y_{22} + y_2)}\right|^2 \tag{14}$$

When y_2 contains the generator and y_1 the load, we may proceed in the same way (letting I_0 flow in terminal 4) and obtain

$$\Gamma_{21} = 4g_1g_2\left|\frac{Y_{12}}{Y_{12}Y_{21} - (Y_{11} + y_1)(Y_{22} + y_2)}\right|^2 \tag{15}$$

We may now obtain expressions for the admittance and gain of the 4-pole when the nonlinear element consists of a nonlinear resistor and a nonlinear capacitor in parallel. We shall do this for the case where a conjugate match exists at the terminals by letting $y_1' = y_1^*$ and $y_2' = y_2^*$. Equations (11) and (12) may thus be written

$$(Y_{11} - y_1^*)(Y_{22} + y_2) = (Y_{11} + y_1)(Y_{22} - y_2^*) = Y_{12}Y_{21} \quad (16)$$

When this is multiplied out, letting $Y_{mn} = G_{mn} + jB_{mn}$, and the real and imaginary parts set equal as indicated by the first equality we obtain $G_{11}g_2 = G_{22}g_1$ and $g_2(B_{11} + b_1) = g_1(B_{22} + b_2)$. In (8) and (9) it is seen that $G_{11} = G_{22} = G_0$ and that B_{22} is positive in equations (8) and negative in equations (9). We thus obtain

$$g_1 = g_2 \qquad b_1 + \omega_1 C_0 = b_2 \pm \omega_2 C_0 \quad (17)$$

where the upper symbol of the $\pm$ sign is used in the noninverting case and the lower symbol in the inverting case. When the real and imaginary parts are set equal as indicated by the second equality in (16) we obtain, using the results in (17),

$$g^2 = G_0^2 - G_1^2 \pm \omega_1\omega_2 C_1^2 - B^2 \quad (18)$$

where

$$g = g_1 = g_2 \quad (19)$$

and

$$B = b_1 + \omega_1 C_0 = b_2 \pm \omega_2 C_0 = \pm \frac{G_1}{2G_0}(\omega_2 \pm \omega_1)C_1 \quad (20)$$

These results may be put in (14) to obtain the modulator gain. Since a conjugate match exists at the terminals of the 4-pole, this is the maximum available gain. The result is

$$MAG_{12} = \frac{G_1^2 + (\omega_2 C_1)^2}{(G_0 + g)^2 + B^2} \quad (21)$$

For the converter, using equation (15) we obtain

$$MAG_{21} = \frac{G_1^2 + (\omega_1 C_1)^2}{(G_0 + g)^2 + B^2} \quad (22)$$

These results are valid only when a conjugate match exists at the terminals. For this to be possible, the right side of (18) must be positive. If it is negative no combination of values of g_1 and g_2 will result in a match.

It may be shown that if the slope of the voltage-current characteristic of the nonlinear resistor is always positive, then G_1/G_0 can never be greater than unity. (Reference 1, p. 410.) It is therefore convenient to normalize the above results with respect to G_0. If we let

$$\frac{\omega_1}{\omega_2} = \rho, \qquad \frac{\omega_1 C_1}{G_0} = \rho x, \qquad \frac{\omega_2 C_1}{G_0} = x, \qquad \frac{G_1}{G_0} = y, \qquad \frac{C_0}{C_1} = z \tag{23}$$

equations (18) through (22) become

$$\left(\frac{g}{G_0}\right)^2 = 1 - y^2 \pm \rho x^2 - \left[(1 \pm \rho)\frac{xy}{2}\right]^2 \tag{24}$$

$$\frac{b_1}{G_0} = \pm (1 \pm p)\frac{xy}{2} - \rho xz, \qquad \frac{b_2}{g_0} = \pm(1 \pm \rho)\frac{xy}{2} \pm xz \tag{25}$$

$$MAG_{12} = \frac{y^2 + x^2}{\left(1 + \frac{g}{G_0}\right)^2 + \left[(1 \pm \rho)\frac{xy}{2}\right]^2} \tag{26}$$

$$MAG_{21} = \frac{y^2 + (\rho x)^2}{\left(1 + \frac{g}{G_0}\right)^2 + \left[(1 \pm \rho)\frac{xy}{2}\right]^2} \tag{27}$$

In these equations, ρ is less than $\frac{1}{3}$ in the noninverting case and less than 1 in the inverting case. Ordinarily it will be very much less than 1. The value of z will be determined by the shape of the nonlinear capacitor characteristic. However z appears only in (25) where it influences the values of the matching susceptances so that it does not affect the conductance or gain. While we can be certain that y will have values between 0 and 1, limitations on the value of x will depend on the particular device used. We will therefore assume that x may have any value.

EFFECT OF NONLINEAR CAPACITOR

We may now examine, in a general way, the manner in which the nonlinear capacitor influences the behavior of the 4-pole. Consider first the case where the nonlinear capacitor is absent. It is well known, and can be seen in the above equations by letting $C_0 = C_1 = 0$, that the noninverting and inverting cases are alike, that the 4-pole can always be matched and that the gain is the same in both directions and can never be greater than unity. In addition, the matching susceptances are zero and the gain is independent of frequency so that there is no limitation to the bandwidth. When the nonlinear capacitor is added, all but one of these conditions are changed. Equations (8) and (9) show that the non-

inverting and inverting cases are different, (24) may become negative so that the 4-pole cannot always be matched and (26) and (27) are different so that the gains through the 4-pole are not the same in the two directions. Furthermore, (26) can be greater than unity so that modulators may have gain. However, as will be shown, the converter gain given by (27) is still restricted to values less than unity. It is also seen that the matching susceptances are no longer zero and that the gain varies with frequency so that the bandwidth is limited.

If we remove the restriction that a conjugate match exists and operate the 4-pole between arbitrary admittances, it may be shown in (11) and (12) that the conductance of the 4-pole may become negative, and in (14) and (15) that the gain may have any value, however large. This is true for both noninverting and inverting modulators and converters. However, we see in (14) and (15) that the ratio of the modulator gain to the converter gain is $|Y_{21}/Y_{12}|^2$. This is greater than unity, so that for the same operating conditions the modulator gain will be greater than the converter gain. Although increased gain is possible, it is obtained at the expense of reduced bandwidth and increased sensitivity to changes in the terminating admittances, particularly in the case of converters. The present analysis will therefore be restricted to the case where a conjugate match exists.

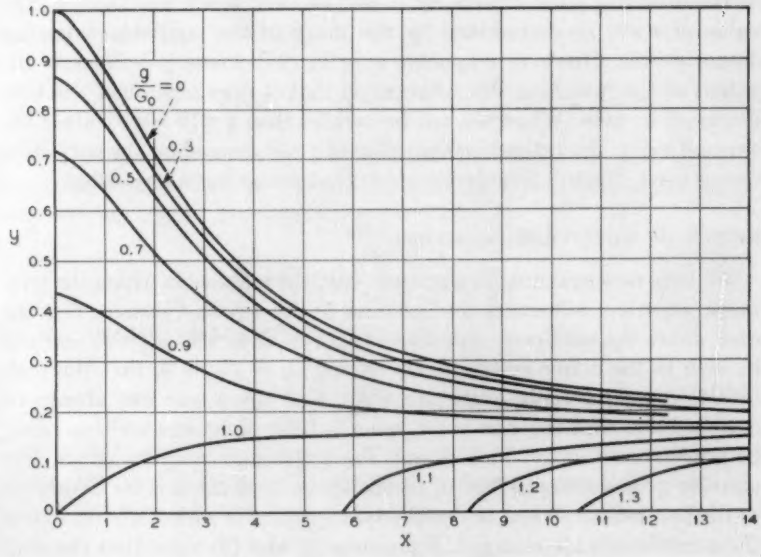


Fig. 2 — Conductance contours of noninverting transducer.

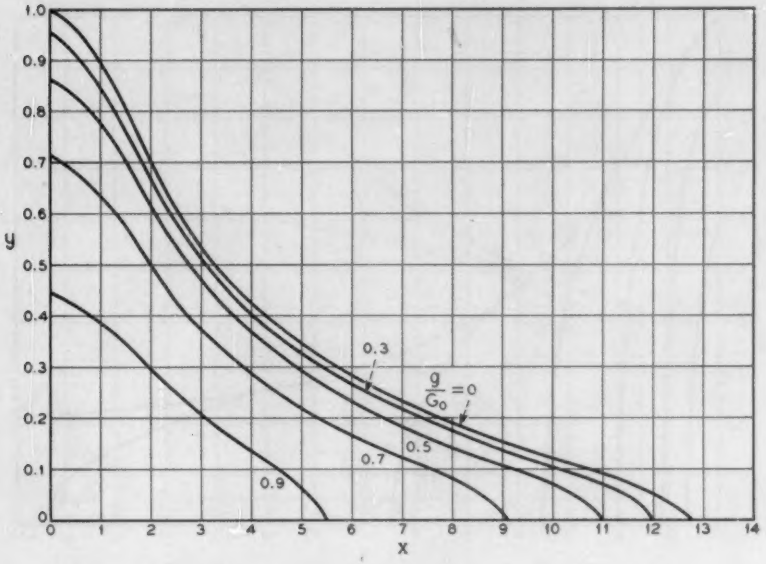


Fig. 3 — Conductance contours for inverting transducer.

CONDUCTANCE AND GAIN VERSUS x AND y

By assigning a value to ρ, curves may be plotted showing how the conductance and gain of the 4-pole change as the characteristics of the nonlinear resistor and nonlinear capacitor are varied. The particular case when f_2 is about 160 times f_1 will be treated. This corresponds, for example, to an intermediate frequency of 70 mc and a local oscillator frequency of 11,200 mc.

Figs. 2 and 3 show the normalized conductance contours as functions of x and y as given by (24) for the noninverting and inverting cases respectively. It will be seen that in most instances, increasing the value of x causes g/G_0 to decrease. An exception occurs in the noninverting case (Fig. 2) when y is less than $2\sqrt{\rho}/(\rho + 1)$ or 0.157 where it is seen that increasing x causes g/G_0 to increase. When x and y have values corresponding to points above the $g/G_0 = 0$ curve, the 4-pole cannot be matched and (23) through (27) are not applicable. However, it will be noted that connecting a resistor across either the nonlinear elements or across the input and output terminals has the effect of increasing G_0. By this means the 4-pole can always be reduced to the condition where it can be matched.

Figs. 4 and 5 show the modulator gain contours as functions of x and y

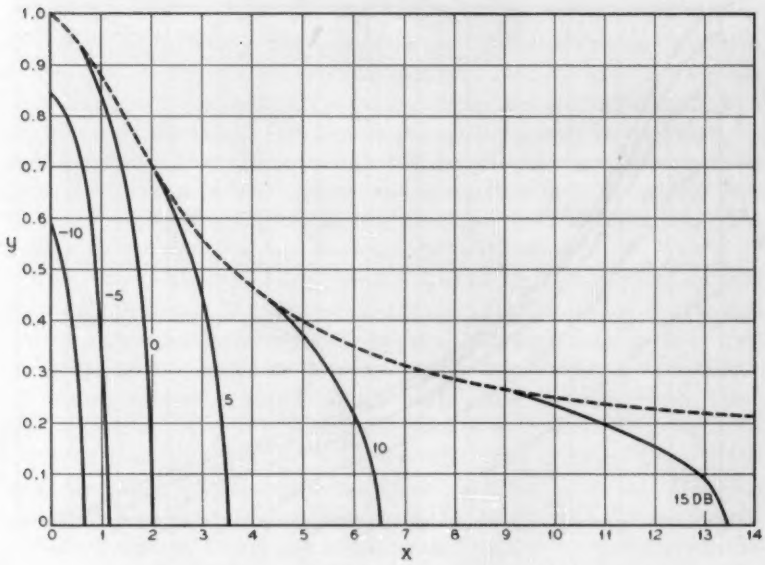


Fig. 4 — Gain contours for noninverting modulators.

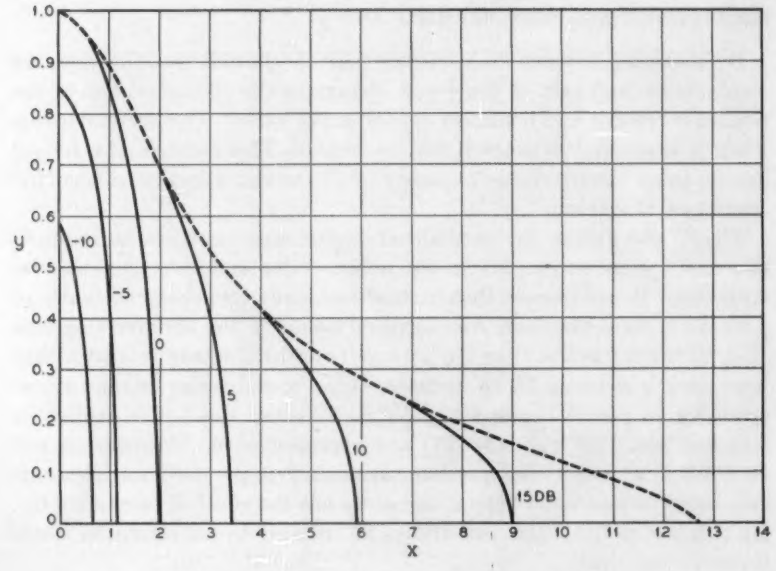


Fig. 5 — Gain contours for inverting modulator.

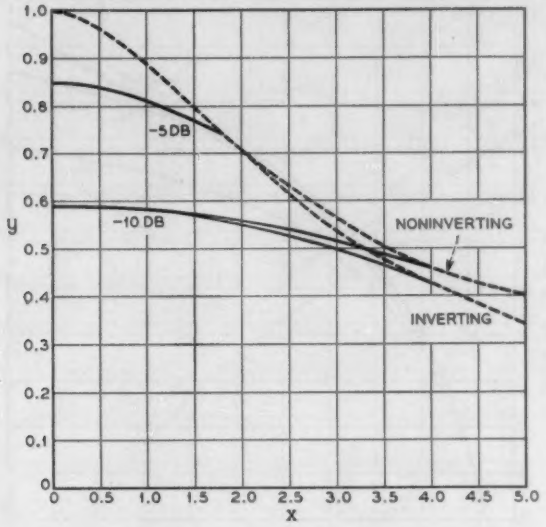


Fig. 6 — Gain contours for converter.

as given by (26). Here it is seen that increasing the value of x causes the gain to increase. For values of x less than about 3, the gains in the noninverting and inverting cases are the same. In the noninverting case, x may increase indefinitely, provided y is less than 0.157, and a gain equal to the ratio of the output frequency to the input frequency eventually reached, 22.1 db in this case. In the inverting case, the maximum gain obtainable is 19.3 db, and it occurs when y is zero.

Fig. 6 shows the converter gain contours as given by equation (27). Here we see that increasing x causes a decrease in the loss, but the decrease is small and in no case can the gain be greater than 0 db. This occurs when x is zero. The nonlinear capacitor is thus of small benefit in the converter case. About the most benefit that can be obtained is a decrease in loss of perhaps 1 db. For example, if the nonlinear resistor alone has a loss of 6 db ($y = 0.8$), this could be reduced to 5 db by adding a nonlinear capacitor of such value as to make $x = 1.3$.

BANDWIDTH

Since both the admittance and gain of the 4-pole vary with frequency, the bandwidth over which it can be used is limited. Figs. 7 and 8 show the modulator gain as a function of x for input frequencies of 50, 70 and 90 mc, and a local oscillator frequency of 11,200 mc. These curves were

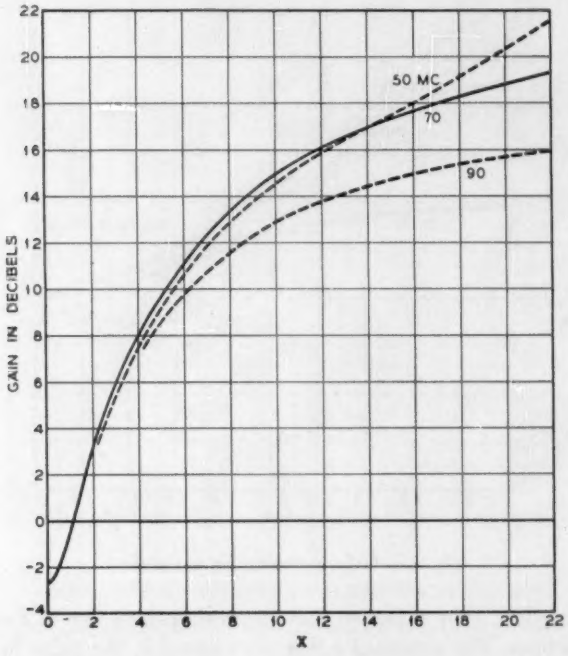


Fig. 7 — Gain of noninverting modulator, $g/G_0 = 0.3$.

computed using values of y which make $g/G_0 = 0.3$ at midband. They are thus near the largest gains obtainable for a given value of x. The matching susceptances were assumed to be a single inductance or capacitance connected across the terminating resistors. C_0/C_1 was arbitrarily assumed to have a value of 2. The procedure used was to compute y, b_1/G_0, b_2/G_0 and the maximum available gain at midband using (24), (25) and (26); b_1/G_0 and b_1/G_0 were then multiplied by the appropriate frequency ratio to obtain the terminating susceptances at 50 and 90 mc and the gain at these frequencies was then computed using (14).

Figs. 7 and 8 show that with the simple matching susceptances used, the gain variation across the band increases as the gain increases. For the same midband gain, the variation in the inverting case is somewhat greater than in the noninverting case. The gain is thus limited by the bandwidth requirements.

When the gain at 50, 70 and 90 mc is calculated using larger values of g/G_0 it is found that as g/G_0 increases the gain variation across the band decreases. In the limit the least variation is obtained when y is

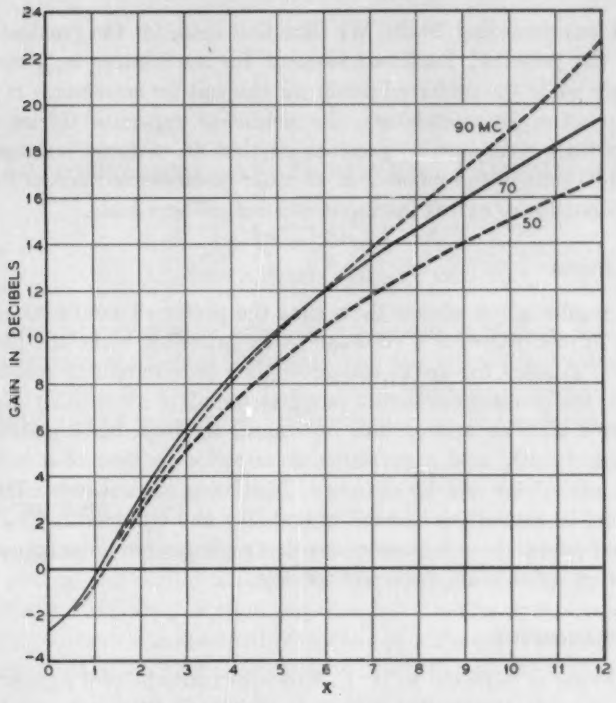


Fig. 8 — Gain of inverting modulator, $g/G_0 = 0.3$.

zero. When the midband gain is 15 db, Figs. 7 and 8 show that the gain variation is 2.0 db in the noninverting case and 2.7 db in the inverting case. When y is zero these variations are reduced to 0.8 db and 1.0 db respectively for the same midband gain. The nonlinear resistor therefore degrades the performance and, assuming complete freedom in the choice of x, a greater bandwidth can be obtained if it is absent.

PREFERRED NONLINEAR ELEMENTS

Thus we see that, under the requirement that a conjugate match exist at the terminals of the 4-pole, the nonlinear resistor contributes little to the gain of a nonlinear capacitor modulator while the nonlinear capacitor is of little benefit in a nonlinear resistor converter. In a modulator having appreciable gain, the degree of nonlinearity permissible in the nonlinear resistor is quite small. For gains exceeding 15 db, y must be less than 0.2. Such a nonlinear resistor used alone would have a con-

version loss exceeding 20 db. We thus find that, for the greatest bandwidth, the preferred nonlinear element for modulators is a nonlinear capacitor while the preferred nonlinear element for converters is a nonlinear resistor. In modulators, the nonlinear capacitor device should have as little resistance as possible, so that an external resistor could be used to control the value of x. It could be connected across the nonlinear capacitor or across the input and output terminals.

CONCLUSIONS

The results given above show that the preferred nonlinear element for use in modulators is a pure nonlinear capacitor while the preferred nonlinear element for use in converters is a pure nonlinear resistor. By shunting the nonlinear capacitor or the terminals of a nonlinear capacitor modulator with an appropriate resistance, an impedance match, adequate bandwidth, and a performance superior to that of a nonlinear resistor modulator can be obtained. Nonlinear capacitance effects are not useful in converters because of stability and bandwidth limitations and also because there is no evidence that an improved noise figure would result from a reduction in conversion loss.

ACKNOWLEDGMENT

The writer is indebted to H. E. Rowe for many helpful suggestions in the mathematical analysis and to R. S. Ohl for supplying the bombarded silicon rectifiers used in the experiments which lead to the ideas presented here.

REFERENCES

1. H. C. Torrey and C. A. Whitmer, Crystal Rectifiers, **15,** Radiation Laboratory Series, McGraw-Hill, New York, 1948, Chapter 13.
2. L. C. Peterson and F. B. Llewellyn, The Performance and Measurement of Mixers in Terms of Linear Network Theory, Proc. I.R.E., **33,** July, 1945.

Minimization of Boolean Functions*

E. J. McCLUSKEY, Jr.

(Manuscript received June 26, 1956)

A systematic procedure is presented for writing a Boolean function as a minimum sum of products. This procedure is a simplification and extension of the method presented by W. V. Quine. Specific attention is given to terms which can be included in the function solely for the designer's convenience.

1 INTRODUCTION

In designing switching circuits such as digital computers, telephone central offices, and digital machine tool controls, it is common practice to make use of Boolean algebra notation.[1, 2, 3, 4] The performance of a single-output circuit is specified by means of a Boolean function of the input variables. This function, which is called the circuit transmission, is equal to 1 when an output is present and equals 0 when there is no output. A convenient means of specifying a transmission is a table of combinations such as that given in Table I. This table lists, in the column under T, the output condition for each combination of input conditions. If there are some combinations of input conditions for which the output is not specified (perhaps because these combinations can never occur), d-entries are placed in the T-column of the corresponding rows of the table of combinations. The actual values (0 or 1) assigned to these rows are usually chosen so as to simplify the circuit which is designed to satisfy the requirements specified in the table of combinations.

For each row of the table of combinations a transmission can be written which equals "one" only when the variables have the values listed in that row of the table. These transmissions will be called *elementary product terms* (or more simply, p-terms) since any transmission can always be written as a sum of these p-terms. Table I (b) lists the p-terms for Table I(a). Note that every variable appears in each p-term. The

* This paper is derived from a thesis submitted to the Massachusetts Institute of Technology in partial fulfillment of the requirements for the degree of Doctor of Science on April 30, 1956.

TABLE I — CIRCUIT SPECIFICATIONS

(a) Table of Combinations

	x_1	x_2	x_3	T
0	0	0	0	0
1	0	0	1	1
2	0	1	0	1
3	0	1	1	1
4	1	0	0	1
5	1	0	1	1
6	1	1	0	1
7	1	1	1	0

(b) p-terms

$x_1' x_2' x_3'$
$x_1' x_2' x_3$
$x_1' x_2 x_3'$
$x_1' x_2 x_3$
$x_1 x_2' x_3'$
$x_1 x_2' x_3$
$x_1 x_2 x_3'$
$x_1 x_2 x_3$

(c) Canonical Expansion

$$T = x_1'x_2'x_3 + x_1'x_2x_3' + x_1'x_2x_3 + x_1x_2'x_3' + x_1x_2'x_3 + x_1x_2x_3'$$

p-term corresponding to a given row of a table of combinations is formed by priming any variables which have a "zero" entry in that row of the table and by leaving unprimed those variables which have "one" entries. It is possible to write an algebraic expression for the over-all circuit transmission directly from the table of combinations. This over-all transmission, T, is the sum of the p-terms corresponding to those rows of the table of combinations for which T is to have the value "one." See Table I(c). Any transmission which is a sum of p-terms is called a *canonical expansion*.

The decimal numbers in the first column of Table I(a) are the decimal equivalents of the binary numbers formed by the entries of the table of combinations. A concise method for specifying a transmission function is to list the decimal numbers of those rows of the table of combinations for which the function is to have the value one. Thus the function of Table I can be specified as $\sum(1, 2, 3, 4, 5, 6)$.

One of the most basic problems of switching circuit theory is that of writing a Boolean function in a simpler form than the canonical expansion. It is frequently possible to realize savings in equipment by writing a circuit transmission in simplified form. Methods for expressing a Boolean function in the "simplest" sum of products form were published by Karnaugh,[1] Aiken,[5] and Quine.[6] These methods have the common property that they all fail when the function to be simplified is reasonably complex. The following sections present a method for simplifying functions which can be applied to more complex functions than previous methods, is systematic, and can be easily programmed on a digital computer.

2 THE MINIMUM SUM

By use of the Boolean algebra theorem $x_1x_2 + x_1'x_2 = x_2$ it is possible to obtain from the canonical expansion other equivalent sum functions:

that is, other sum functions which correspond to the same table of combinations. These functions are still sums of products of variables but not all of the variables appear in each term. For example, the transmission of Table I, $T = x_1'x_2'x_3 + x_1'x_2x_3' + x_1'x_2x_3 + x_1x_2'x_3' + x_1x_2'x_3 + x_1x_2x_3' = (x_1'x_2'x_3 + x_1'x_2x_3) + (x_1'x_2x_3' + x_1x_2x_3') + (x_1x_2'x_3' + x_1x_2'x_3) = (x_1'x_2'x_3 + x_1x_2'x_3) + (x_1'x_2x_3' + x_1'x_2x_3) + (x_1x_2'x_3' + x_1x_2x_3')$ can be written as either $T = x_1'x_3 + x_2x_3' + x_1x_2'$ or $T = x_2'x_3 + x_1'x_2 + x_1x_3'$.

A *literal* is defined as a variable with or without the associated prime (x_1 , x_2' are literals). The sum functions which have the fewest terms of all equivalent sum functions will be called *minimum sums* unless these functions having fewest terms do not all involve the same number of literals. In such cases, only those functions which involve the fewest literals will be called minimum sums. For example, the function

$$T = \sum(7, 9, 10, 12, 13, 14, 15)$$

can be written as either

$$T = x_4x_2x_1' + x_3x_2x_1 + x_4x_2'x_1 + x_4x_3x_1'$$

or as

$$T = x_4x_2x_1' + x_3x_2x_1 + x_4x_2'x_1 + x_4x_3$$

Only the second expression is a minimum sum since it involves 11 literals while the first expression involves 12 literals.

The minimum sum defined here is not necessarily the expression containing the fewest total literals, or the expression leading to the most economical two-stage diode logic circuit,[1] even though these three expressions are identical for many transmissions. The definition adopted here lends itself well to computation and results in a form which is useful in the design of contact networks. A method is presented in Section 9 for obtaining directly the expressions corresponding to the optimum two-stage diode logic circuit or the expressions containing fewest literals.

In principle it is possible to obtain a minimum sum for any given transmission by enumerating all possible equivalent sum functions then selecting those functions which have the fewest terms, and finally selecting from these the functions which contain fewest literals. Since the number of equivalent sum functions may be quite large, this procedure is not generally practical. The following sections present a practical method for obtaining a minimum sum without resorting to an enumeration of all equivalent sum functions.

3 PRIME IMPLICANTS

When the theorem $x_1x_2 + x_1x_2' = x_1$ is used to replace by a single term, two p-terms, which correspond to rows i and j of a table of combi-

nations, the resulting term will equal "one" when the variables have values corresponding to either row i or row j of the table. Similarly, when this theorem is used to replace, by a single term, a term which equals "one" for rows i and j and a term which equals "one" for rows k and m, the resulting term will equal "one" for rows i, j, k and m of the table of combinations. A method for obtaining a minimum sum by repeated application of this theorem ($x_1x_2' + x_1x_2 = x_1$) was first presented by Quine.[6] In this method, the theorem is applied to all possible pairs of p-terms, then to all possible pairs of the terms obtained from the p-terms, and so on, until no further applications of the theorem are possible. It may be necessary to pair one term with several other terms in applying this theorem. In Example 3.2 the theorem is applied to the terms labeled 5 and 7 and also to the terms labeled 5 and 13. All terms paired with other terms in applying the theorem are then discarded. The remaining terms are called *prime implicants.*[6] Finally a minimum sum is formed as the sum of the fewest prime implicants which when taken together will equal "one" for all required rows of the table of combinations. The terms in the minimum sum will be called *minimum sum terms or ms-terms.*

Example 3.1

$$T = \sum(3, 7, 8, 9, 12, 13)$$

Canonical Expansion:

$$T = \underset{\begin{bmatrix}0\ 0\ 1\ 1\\ 3\end{bmatrix}}{x_1'x_2'x_3x_4} + \underset{\begin{bmatrix}0\ 1\ 1\ 1\\ 7\end{bmatrix}}{x_1'x_2x_3x_4} + \underset{\begin{bmatrix}1\ 0\ 0\ 0\\ 8\end{bmatrix}}{x_1x_2'x_3'x_4'} + \underset{\begin{bmatrix}1\ 0\ 0\ 1\\ 9\end{bmatrix}}{x_1x_2'x_3'x_4} + \underset{\begin{bmatrix}1\ 1\ 0\ 0\\ 12\end{bmatrix}}{x_1x_2x_3'x_4'} + \underset{\begin{bmatrix}1\ 1\ 0\ 1\\ 13\end{bmatrix}}{x_1x_2x_3'x_4}$$

The bracketed binary and decimal numbers below the sum terms indicate the rows of the table of combinations for which the corresponding term will equal "one." A binary character in which a dash appears represents the two binary numbers which are formed by replacing the dash by a "0" and then by a "1." Similarly a binary character in which two dashes appear represents the four binary numbers formed by replacing the dashes by "0" and "1" entries, etc.

$$\underset{\begin{bmatrix}0\ 0\ 1\ 1\\ 3\end{bmatrix}}{x_1'x_2'x_3x_4} + \underset{\begin{bmatrix}0\ 1\ 1\ 1\\ 7\end{bmatrix}}{x_1'x_2x_3x_4} = \underset{\begin{bmatrix}0-1\ 1\\ 3,7\end{bmatrix}}{x_1'\ x_3x_4}$$

$$x_1x_2'x_3'x_4' + x_1x_2'x_3'x_4 = x_1x_2'x_3'$$
$$\begin{bmatrix}1\,0\,0\,0\\8\end{bmatrix}\quad\begin{bmatrix}1\,0\,0\,1\\9\end{bmatrix}\quad\begin{bmatrix}1\,0\,0\,-\\8,9\end{bmatrix}$$

$$x_1x_2x_3'x_4' + x_1x_2x_3'x_4 = x_1x_2x_3'$$
$$\begin{bmatrix}1\,1\,0\,0\\12\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,1\\13\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,-\\12,13\end{bmatrix}$$

$$x_1x_2'x_3' + x_1x_2x_3' = x_1\ x_3'$$
$$\begin{bmatrix}1\,0\,0\,-\\8,9\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,-\\12,13\end{bmatrix}\quad\begin{bmatrix}1-0\,-\\8,9,12,13\end{bmatrix}$$

Prime Implicants:

$$x_1\ x_3',\qquad x_1'\ x_2x_4$$
$$\begin{bmatrix}1-0\,-\\8,9,12,13\end{bmatrix}\begin{bmatrix}0-1\,1\\3,7\end{bmatrix}$$

Minimum Sum:

$$T = x_1x_3' + x_1'x_2x_4$$

Example 3.2

$$T = \sum(5, 7, 12, 13)$$

Canonical Expansion:

$$T = x_1'x_2x_3'x_4 + x_1'x_2x_3x_4 + x_1x_2x_3'x_4' + x_1x_2x_3'x_4$$
$$\begin{bmatrix}0\,1\,0\,1\\5\end{bmatrix}\quad\begin{bmatrix}0\,1\,1\,1\\7\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,0\\12\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,1\\13\end{bmatrix}$$

$$x_1'x_2x_3'x_4 + x_1'x_2x_3x_4 = x_1'x_2\ \ x_4$$
$$\begin{bmatrix}0\,1\,0\,1\\5\end{bmatrix}\quad\begin{bmatrix}0\,1\,1\,1\\7\end{bmatrix}\quad\begin{bmatrix}0\,1-1\\5,7\end{bmatrix}$$

$$x_1'x_2x_3'x_4 + x_1x_2x_3'x_4 = x_2x_3'x_4$$
$$\begin{bmatrix}0\,1\,0\,1\\5\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,1\\13\end{bmatrix}\quad\begin{bmatrix}-1\,0\,1\\5,13\end{bmatrix}$$

$$x_1x_2x_3'x_4' + x_1x_2x_3'x_4 = x_1x_2x_3'$$
$$\begin{bmatrix}1\,1\,0\,0\\12\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,1\\13\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,-\\12,13\end{bmatrix}$$

Prime Implicants:

$$x_1'x_2\ \ x_4,\qquad x_2x_3'x_4,\qquad x_1x_2x_3'$$
$$\begin{bmatrix}0\,1-1\\5,7\end{bmatrix}\quad\begin{bmatrix}-1\,0\,1\\5,13\end{bmatrix}\quad\begin{bmatrix}1\,1\,0\,-\\12,13\end{bmatrix}$$

Minimum Sum:

$$T = x_1'x_2x_4 + x_1x_2x_3'$$

Quine's method, as illustrated in Examples 3.1 and 3.2, becomes unwieldly for transmissions involving either many variables or many p-terms. This difficulty is overcome by simplifying the notation and making the procedure more systematic. The notation is simplified by discarding the expressions involving literals and using only the binary characters. This is permissible because the expressions in terms of literals can always be regained from the binary characters. The theorem being used to combine terms can be stated in terms of the binary characters as follows: If two binary characters are identical in all positions except one, and if neither character has a dash in the position in which they differ, then the two characters can be replaced by a single character which has a dash in the position in which they differ and which is identical with the original characters in all other positions.

TABLE II — DETERMINATION OF PRIME IMPLICANTS FOR TRANSMISSION

$$T = \sum (0, 2, 4, 6, 7, 8, 10, 11, 12, 13, 14, 16, 18, 19, 29, 30)$$

(a) I

	$x_5x_4x_3x_2x_1$	
0	0 0 0 0 0	√
2	0 0 0 1 0	√
4	0 0 1 0 0	√
8	0 1 0 0 0	√
16	1 0 0 0 0	√
6	0 0 1 1 0	√
10	0 1 0 1 0	√
12	0 1 1 0 0	√
18	1 0 0 1 0	√
7	0 0 1 1 1	√
11	0 1 0 1 1	√
13	0 1 1 0 1	√
14	0 1 1 1 0	√
19	1 0 0 1 1	√
29	1 1 1 0 1	√
30	1 1 1 1 0	√

(b) II

	$x_5x_4x_3x_2x_1$	
0 2	0 0 0 – 0	√
0 4	0 0 – 0 0	√
0 8	0 – 0 0 0	√
0 16	– 0 0 0 0	√
2 6	0 0 – 1 0	√
2 10	0 – 0 1 0	√
2 18	– 0 0 1 0	√
4 6	0 0 1 – 0	√
4 12	0 – 1 0 0	√
8 10	0 1 0 – 0	√
8 12	0 1 – 0 0	√
16 18	1 0 0 – 0	√
6 7	0 0 1 1 –	
6 14	0 – 1 1 0	√
10 11	0 1 0 1 –	
10 14	0 1 – 1 0	√
12 13	0 1 1 0 –	
12 14	0 1 1 – 0	√
18 19	1 0 0 1 –	
13 29	– 1 1 0 1	
14 30	– 1 1 1 0	

(c) III

	$x_5x_4x_3x_2x_1$	
0 2 4 6	0 0 – – 0	√
0 2 8 10	0 – 0 – 0	√
0 2 16 18	– 0 0 – 0	
0 4 8 12	0 – – 0 0	√
2 6 10 14	0 – – 1 0	√
4 6 12 14	0 – 1 – 0	√
8 10 12 14	0 1 – – 0	√

(d) IV

	$x_5x_4x_3x_2x_1$
0 2 4 6 8 10 12 14	0 – – – 0

The first step in the revised method for determining prime implicants is to list in a column, such as that shown in Table II(a), the binary equivalents of the decimal numbers which specify the function. It is expedient to order these binary numbers so that any numbers which contain no 1's come first, followed by any numbers containing a single 1, etc. Lines should be drawn to divide the column into groups of binary numbers which contain a given number of 1's. The theorem stated above is applied to these binary numbers by comparing each number with all the numbers of the next lower group. Other pairs of numbers need not be considered since any two numbers which are not from adjacent groups must differ in more than one binary digit. For each number which has 1's wherever the number (from the next upper group) with which it is being compared has 1's, a new character is formed according to the theorem. A check mark is placed next to each number which is used in forming a new character. The new characters are placed in a separate column, such as Table II(b), which is again divided into groups of characters which have the same number of 1's. The characters in this new column will each contain one dash.

After each number in the first column has been considered, a similar process is carried out for the characters of column two. Two characters from adjacent groups can be combined if they both have their dashes in the same position and if the character from the lower group has 1's wherever the upper character has 1's. If any combinations are possible the resulting characters are placed in a third column such as Table II(c), and the Column II characters from which the new characters are formed are checked. All the characters in this third column will have two dashes. This procedure is repeated and new columns are formed, Table II(d), until no further combinations are possible. The unchecked characters, which have not entered into any combinations, represent the prime implicants.

Each binary character is labeled with the decimal equivalents of the binary numbers which it represents (see note in Example 3.1). These decimal numbers are arranged in increasing arithmetic order. For a character having one dash this corresponds to the order of its formation: When two binary numbers combine, the second number always contains all the 1's of the first number and one additional 1 so that the second number is always greater than the first. Characters having two dashes can be formed in two ways. For example, the character (0, 2, 4, 6) can be formed either by combining (0, 2) and (4, 6) or by combining (0, 4) and (2, 6) as given in Table III. Similarly, there are three ways in which a character having three dashes can be formed (in Table II the 0, 2, 4,

TABLE III — EXAMPLE OF THE TWO WAYS OF FORMING A CHARACTER HAVING TWO DASHES

0	0 0 0 0	0 2	0 0 – 0	0 2 4 6	0 – – 0
		0 4	0 – 0 0	(0 4 2 6	0 – – 0)
2	0 0 1 0				
4	0 1 0 0	2 6	0 – 1 0		
		4 6	0 1 – 0		
6	0 1 1 0				

6, 8, 10, 12, 14 character can be formed from the 0, 2, 4, 6, and 8, 10, 12, 14 characters or the 0, 2, 8, 10, and 4, 6, 12, 14 characters or the 0, 4, 8, 12 and 2, 6, 10, 14 characters), four ways in which a character having four dashes can be formed, etc.

In general, any character can be formed by combining two characters whose labels form an increasing sequence of decimal numbers when placed together. It is possible to shorten the process of determining prime implicants by not considering the combination of any characters whose labels do not satisfy this requirement. For example, in Table II(b) the possibility of combining the (0, 4) character with either the (2, 6), (2, 10) or the (2, 18) character need not be considered. If the process is so shortened, it is not sufficient to place check marks next to the two characters from which a new character is formed; each member of all pairs of characters which would produce the same new character when combined must also receive check marks. More simply, when a new character is formed a check mark is placed next to all characters whose labels contain only decimal numbers which occur in the label of the new character. In Table II, when the (0, 2, 4, 6) character is formed by combining the (0, 2) and (4, 6) characters, check marks must be placed next to the (0, 4) and (2, 6) characters as well as the (0, 2) and (4, 6) characters. If the process is not shortened as just described, the fact that a character can be formed in several ways can serve as a check on the accuracy of the process.

It is possible to carry out the entire process of determining the prime implicants solely in terms of the decimal labels without actually writing the binary characters. If two binary characters can be combined as described in this section, then the decimal label of one can be obtained from the decimal label of the other character by adding some power of two (corresponding to the position in which the two characters differ) to each number in the character's label. For example, in Table IIb the label of the (4, 6)(0 0 1 – 0) character can be obtained by adding $4 = (2^2)$ to the numbers of the label of the (0, 2)(0 0 0 – 0) character. By searching for decimal labels which differ by a power of two, instead of binary characters which differ in only one position, the prime implicants can be

determined as described above without ever actually writing the binary characters.

4 PRIME IMPLICANT TABLES

The minimum sum is formed by picking the fewest prime implicants whose sum will equal one for all rows of the table of combinations for which the transmission is to equal one. In terms of the characters used in Section 3 this means that each number in the decimal specification of the function must appear in the label of at least one character which corresponds to a ms-term (term of the minimum sum).

The ms-terms are selected from the prime implicants by means of a prime implicant table,* Table IV. Each column of the prime implicant table corresponds to a row of the table of combinations for which the transmission is to have the value one. The decimal number at the top of each column specifies the corresponding row of the table of combinations. Thus the numbers which appear at the tops of the columns are the same as those which specify the transmission. Each row of the prime implicant table represents a prime implicant. If a prime implicant equals "one" for a given row of the table of combinations, a cross is placed at the intersection of the corresponding row and column of the prime implicant table. All other positions are left blank. The table can be written directly from the characters obtained in Section 3 by identifying each row of the table with a character and then placing a cross in each column whose number appears in the label of the character.

It is convenient to arrange the rows in the order of the number of crosses they contain, with those rows containing the most crosses at the top of the table. Also, horizontal lines should be drawn partitioning the table into groups of rows which contain the same number of crosses, Table IV. If, in selecting the rows which are to correspond to ms-terms, a choice between two equally appropriate rows is required, the row having more crosses should be selected. The row with more crosses has fewer literals in the corresponding prime implicant. This choice is more obvious when the table is partitioned as suggested above.

A minimum sum is determined from the prime implicant table by selecting the fewest rows such that each column has a cross in at least one selected row. The selected rows are called *basis rows*, and the prime implicants corresponding to the basis rows are the ms-terms. If any column has only one entry, the row in which this entry occurs must be a basis row. Therefore the first step in selecting the basis rows is to place

* This table was first discussed by Quine.[4] However, no systematic procedure for obtaining a minimum sum from the prime implicant table was presented.

TABLE IV — PRIME IMPLICANT TABLE FOR THE TRANSMISSION OF TABLE II

	0	2	4	8	16	6	10	12	18	7	11	13	14	19	29	30	
A	x	x	x	x		x	x	x					x				*
B	x	x			x				x								*
C													x			x	*
D												x			x		*
E									x					x			*
F								x				x					*
G							x				x						*
H						x				x							*

an asterisk next to each row which contains the sole entry of any column (rows A, B, C, D, E, G, H, in Table IV). A line is then drawn through all rows marked with an asterisk and through *all* columns in which these rows have entries. This is done because the requirement that these columns have entries in at least one basis row is satisfied by selecting the rows marked with an asterisk as basis rows. When this is done for Table IV, all columns are lined out and therefore the rows marked with asterisks are the basis rows for this table. Since no alternative choice of basis rows is possible, there is only one minimum sum for the transmission described in this table.

5 ROW COVERING

In general, after the appropriate rows have been marked with asterisks and the corresponding columns have been lined out, there may remain some columns which are not lined out; for example, column 7 in Table V(b). When this happens, additional rows must be selected and the columns in which these rows have entries must be lined out until all columns of the table are lined out. For Table V(b), the selection of either row B or row F as a basis row will cause column 7 to be lined out. However, row B is the correct choice since it has more crosses than row F. This is an example of the situation which was described earlier in connection with the partitioning of prime implicant tables. Row B is marked with two asterisks to indicate that it is a basis row even though it does not contain the sole entry of any column.

The choice of basis rows to supplement the single asterisk rows becomes more complicated when several columns (such as columns 2, 3, and 6 in Table VI(a)) remain to be lined out. The first step in choosing these supplementary basis rows is to determine whether any pairs of rows exist such that one row has crosses only in columns in which the

TABLE V — DETERMINATION OF THE MINIMUM SUM FOR

$$T = \sum (0, 1, 2, 3, 7, 14, 15, 22, 23, 29, 31)$$

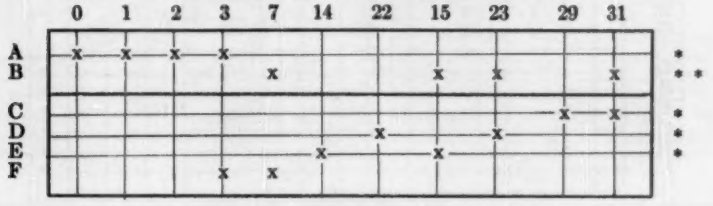
(a) Determination of Prime Implicants

	$x_5x_4x_3x_2x_1$
0	0 0 0 0 0 √
1	0 0 0 0 1 √
2	0 0 0 1 0 √
3	0 0 0 1 1 √
7	0 0 1 1 1 √
14	0 1 1 1 0 √
22	1 0 1 1 0 √
15	0 1 1 1 1 √
23	1 0 1 1 1 √
29	1 1 1 0 1 √
31	1 1 1 1 1 √

	$x_5x_4x_3x_2x_1$
0 1	0 0 0 0 - √
0 2	0 0 0 - 0 √
1 3	0 0 0 - 1 √
2 3	0 0 0 1 - √
3 7	0 0 - 1 1
7 15	0 - 1 1 1 √
7 23	- 0 1 1 1 √
14 15	0 1 1 1 -
22 23	1 0 1 1 -
15 31	- 1 1 1 1 √
23 31	1 - 1 1 1 √
29 31	1 1 1 - 1

	$x_5x_4x_3x_2x_1$
0 1 2 3 7 15 23 31	0 0 0 - -
	- - 1 1 1

(b) First Step in Selection of Basis Rows



(c) Minimum Sum

$$T = \Sigma\,[(0, 1, 2, 3), (7, 15, 23, 31), (29, 31), (22, 23), (14, 15)]$$

$$T = x_5'x_4'x_3' + x_3x_2x_1 + x_5x_4x_3x_1 + x_5x_4'x_3x_2 + x_5'x_4x_3x_2$$

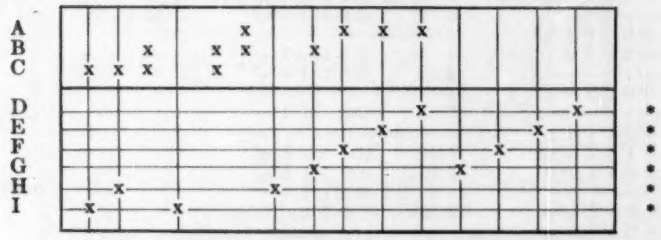
other member of the pair has crosses. Crosses in lined-out columns are not considered. In Table VI(a), rows A and B and rows B and C are such pairs of rows since row B has crosses in columns 2, 3, and 6 and row A has a cross in column 6 and row C has crosses in columns 2 and 3. A convenient way to describe this situation is to say that row B *covers* rows A and C, and to write B $\supset$ A, B $\supset$ C. If row i is selected as a supplementary basis row and row i is covered by row j, which has the same total number of crosses as row i, then it is possible to choose row j as a basis row instead of row i since row j has a cross in each column in which row i has a cross.

The next step is to line out any *rows* which are covered by other rows in the same partition of the table, rows A and C in Table VI(a). If any

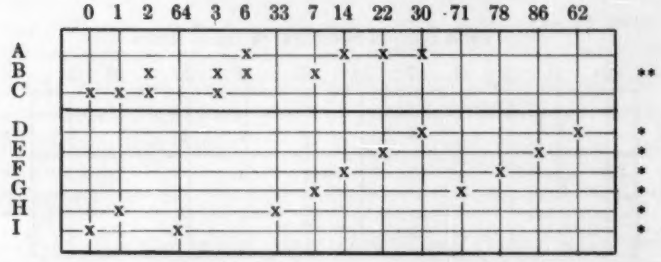
TABLE VI — PRIME IMPLICANT TABLES FOR

$T = \sum (0, 1, 2, 3, 6, 7, 14, 22, 30, 33, 62, 64, 71, 78, 86)$

(a) Prime Implicant Table with Single Asterisk Rows and Corresponding Columns Lined Out



(b) Prime Implicant Table with Rows which are Covered by Other Rows Lined Out



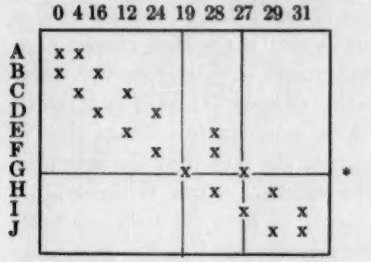
column now contains only one cross which is not lined out, columns 2 3, and 6 in Table VI(b), two asterisks are placed next to the row in which this cross occurs, row B in Table VI(b), and this row and *all* columns in which this row has crosses are lined out. The process of drawing a line through any row which is covered by another row and selecting each row which contains the only cross in a column is continued until it terminates. Either all columns will be lined out, in which case the rows marked with one or two asterisks are the basis rows, or each column will contain more than one cross and no row will cover another row. The latter situation is discussed in the following section.

6 PRIME IMPLICANT TABLES IN CYCLIC FORM

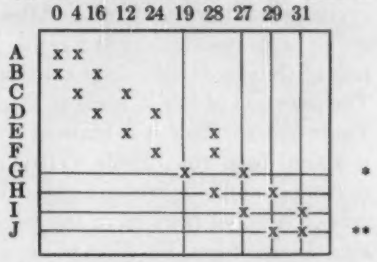
If the rows and columns of a table which are not lined out are such that every column has more than one cross and no row covers another row, as in Table VII(b), the table will be said to be in *cyclic form*, or, in short,

TABLE VII — DETERMINATION OF BASIS ROWS FOR A CYCLIC PRIME IMPLICANT TABLE

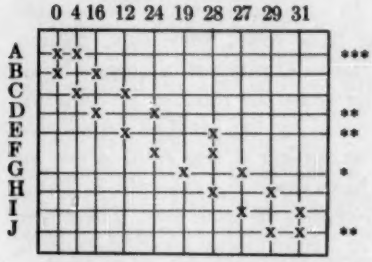
(a) Selection of Single Asterisk Rows



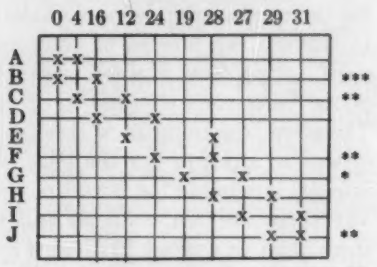
(b) Selection of Double Asterisk Rows



(c) Selection of Row 1 as a Trial Basis Row (Column 0)



(d) Selection of Row 2 as a Trial Basis Row (Column 0)



to be *cyclic*. If any column has crosses in only two rows, at least one of these rows must be included in any set of basis rows. Therefore, the basis rows for a cyclic table can be discovered by first determining whether any column contains only two crosses, and if such a column exists, by then selecting as a trial basis row one of the rows in which the crosses of this column occur. If no column contains only two crosses, then a column which contains three crosses is selected, etc. *All* columns in which the trial basis row has crosses are lined out and the process of lining out rows which are covered by other rows and selecting each row which contains the only cross of some column is carried out as described above. Either all columns will be lined out or another cyclic table will result. Whenever a cyclic table occurs, another trial row must be selected. Eventually all columns will be lined out. However, there is no guarantee that the selected rows are actually basis rows. The possibility exists that a different choice of trial rows would have resulted in fewer selected rows. In general, it is necessary to carry out the procedure of selecting rows several times, choosing different trial rows each time, so

that all possible combinations of trial rows are considered. The set of fewest selected rows is the actual set of basis rows.

Table VII illustrates the process of determining basis rows for a cyclic prime implicant table. After rows G and J have been selected a cyclic table results, Table VII(b). Rows A and B are then chosen as a pair of trial basis rows since column 0 has crosses in only these two rows. The selection of row A leads to the selection of rows D and E as given in Table VII(c). Row A is marked with three asterisks to indicate that it is a trial basis row. Table VII(d) illustrates the fact that the selection of rows C and F is brought about by the selection of row B. Since both sets of selected rows have the same number of rows (5) they are both sets of basis rows. Each set of basis rows corresponds to a different minimum sum so that there are two minimum sums for this function.

Sometimes it is not necessary to determine all minimum sums for the transmission being considered. In such cases, it may be possible to shorten the process of determining basis rows. Since each column must have a cross in some basis row, the total number of crosses in all of the basis rows is equal to or greater than the number of columns. Therefore, the number of columns divided by the greatest number of crosses in any row (or the next highest integer if this ratio is not an integer) is equal to the fewest possible basis rows. For example, in Table VII there are ten columns and two crosses in each row. Therefore, there must be at least 10 divided by 2 or 5 rows in any set of basis rows. The fact that there are only five rows selected in Table VII(c) guarantees that the selected rows are basis rows and therefore Table VII(d) is unnecessary if only one minimum sum is required. In general, the process of trying different combinations of trial rows can be stopped as soon as a set of selected rows which contains the fewest possible number of basis rows has been found (providing that it is not necessary to discover *all* minimum sums). It should be pointed out that more than the minimum number of basis rows may be required in some cases and in these cases all combinations of trial rows must be considered. A more accurate lower bound on the number of basis rows can be obtained by considering the number of rows which have the most crosses. For example, in Table VI there are 15 columns and 4 crosses, at most, in any row. A lower bound of 4 ($\frac{15}{4} = 3\frac{3}{4}$) is a little too optimistic since there are only three rows which contain four crosses. A more realistic lower bound of 5 is obtained by noting that the rows which have 4 crosses can provide crosses in at most 12 columns and that at least two additional rows containing two crosses are necessary to provide crosses in the three remaining columns.

7 CYCLIC PRIME IMPLICANT TABLES AND GROUP INVARIANCE

It is not always necessary to resort to enumeration in order to determine all minimum sums for a cyclic prime implicant table. Often there is a simple relation among the various minimum sums for a transmission so that they can all be determined directly from any single minimum sum by simple interchanges of variables. The process of selecting basis rows for a cyclic table can be shortened by detecting beforehand that the minimum sums are so related.

An example of a transmission for which this is true is given in Table VIII. If the variables x_1 and x_2 are interchanged, one of the minimum sums is changed into the other. In the prime implicant table the interchange of x_1 and x_2 leads to the interchange of columns 1 and 2, 5 and 6, 9 and 10, 13 and 14, and rows A and B, C and D, E and F, G and H. The transmission itself remains the same after the interchange.

In determining the basis rows for the prime implicant table, Table VIII(d), either row G or row H can be chosen as a trial basis row. If row G is selected the i-set of basis rows will result and if row H is selected the ii-set of basis rows will result. It is unnecessary to carry out the procedure of determining both sets of basis rows. Once the i-set of basis rows is known, the ii-set can be determined directly by interchanging the x_1 and x_2 variables in the i-set. Thus no enumeration is necessary in order to determine all minimum sums.

In general, the procedure for a complex prime implicant table is to determine whether there are any pairs of variables which can be interchanged without effecting the transmission. If such pairs of variables exist, the corresponding interchanges of pairs of rows are determined. A trial basis row is then selected from a pair of rows which contain the only two crosses of a column and which are interchanged when the variables are permuted. After the set of basis rows has been determined, the other set of basis rows can be obtained by replacing each basis row by the row with which it is interchanged when variables are permuted. If any step of this procedure is not possible, it is necessary to resort to enumeration.

In the preceding discussion only simple interchanges of variables have been mentioned. Actually all possible permutations of the contact variables should be considered. It is also possible that priming variables or both priming and permuting them will leave the transmission unchanged. For example, if $T = x_4 x_3' x_2 x_1' + x_4' x_3 x_2' x_1$, priming all the variables leaves the function unchanged. Also, priming x_4 and x_3 and then interchanging x_4 and x_3 does not change the transmission. The general name for this property is *group invariance*. This was discussed by Shannon.[4]

A method for determining the group invariance for a specified transmission is presented in "Detection of Group Invariance or Total Symmetry of a Boolean Function."*

8 AN APPROXIMATE SOLUTION FOR CYCLIC PRIME IMPLICANT TABLES

It has not been possible to prove in general that the procedure presented in this section will always result in a minimum sum. However, this procedure should be useful when a reasonable approximation to a minimum sum is sufficient, or when it is possible to devise a proof to show that the procedure does lead to a minimum sum for a *specific transmission* (such proofs were discussed in Section 6). Since this procedure is much simpler than enumeration, it should generally be tested before resorting to enumeration.

The first step of the procedure is to select from the prime implicant table a set of rows such that (1) in each column of the table there is a cross from at least one of the selected rows and (2) none of the selected rows can be discarded without destroying property (1). Any set of rows having these properties will be called a *consistent row set*. Each consistent row set corresponds to a sum of products expression from which no product term can be eliminated directly by any of the theorems of Boolean Algebra. In particular, the consistent row sets having the fewest members correspond to minimum sums. The first step of the procedure to be described here is to select a consistent row-set. This is done by choosing one of the columns, counting the total number of crosses in each row which has a cross in this column, and then selecting the row with the most crosses. If there is more than one such row, the topmost row is arbitrarily selected. The selected row is marked with a check. In Table IX, column 30 was chosen and then row A was selected since rows A and Z each have a cross in column 30, but row A has 4 crosses while row Z has only 2 crosses. The selected row and each column in which it has a cross is then lined out. The process just described is repeated by selecting another column (which is not lined out). Crosses in lined-out columns are not counted in determining the total number of crosses in a row. The procedure is repeated until all columns are lined out.

The table is now rearranged so that all of the selected rows are at the top, and a line is drawn to separate the selected rows from the rest. Table X results from always choosing the rightmost column in Table IX. If any column contains only one cross from a selected row, the single selected-row cross is circled. Any selected row which does not have any

* See page 1445 of this issue.

TABLE VIII — DETERMINATION OF THE MINIMUM SUMS FOR
$T = \sum (0, 1, 2, 5, 6, 7, 9, 10, 11, 13, 14, 15)$

(a)

	$x_4x_3x_2x_1$	
0	0 0 0 0	√
1	0 0 0 1	√
2	0 0 1 0	√
5	0 1 0 1	√
6	0 1 1 0	√
9	1 0 0 1	√
10	1 0 1 0	√
7	0 1 1 1	√
11	1 0 1 1	√
13	1 1 0 1	√
14	1 1 1 0	√
15	1 1 1 1	√

(b)

	$x_4x_3x_2x_1$	
0, 1	0 0 0 –	
0, 2	0 0 – 0	
1, 5	0 – 0 1	√
1, 9	– 0 0 1	√
2, 6	0 – 1 0	√
2, 10	– 0 1 0	√
5, 7	0 1 – 1	√
5, 13	– 1 0 1	√
6, 7	0 1 1 –	√
6, 14	– 1 1 0	√
9, 11	1 0 – 1	√
9, 13	1 – 0 1	√
10, 11	1 0 1 –	√
10, 14	1 – 1 0	√
7, 15	– 1 1 1	√
11, 15	1 – 1 1	√
13, 15	1 1 – 1	√
14, 15	1 1 1 –	√

(c)

	$x_4x_3x_2x_1$
1 5 9 13	– – 0 1
2 6 10 14	– – 1 0
5 7 13 15	– 1 – 1
6 7 14 15	– 1 1 –
9 11 13 15	1 – – 1
10 11 14 15	1 – 1 –

(d)

	0	1	2	5	6	9	10	7	11	13	14	15
A		x		x		x				x		
B			x		x		x				x	
C				x				x		x		x
D					x			x			x	x
E						x			x	x		x
F							x		x		x	x
G	x	x										
H	x		x									

(e)

(i) (0, 1) + (2, 6, 10, 14) + (5, 7, 13, 15) + (9, 11, 13, 15)
(ii) (0, 2) + (1, 5, 9, 13) + (6, 7, 14, 15) + (10, 11, 14, 15)

$$T_i = x_4'x_3'x_2' + x_2x_1' + x_3x_1 + x_4x_1$$

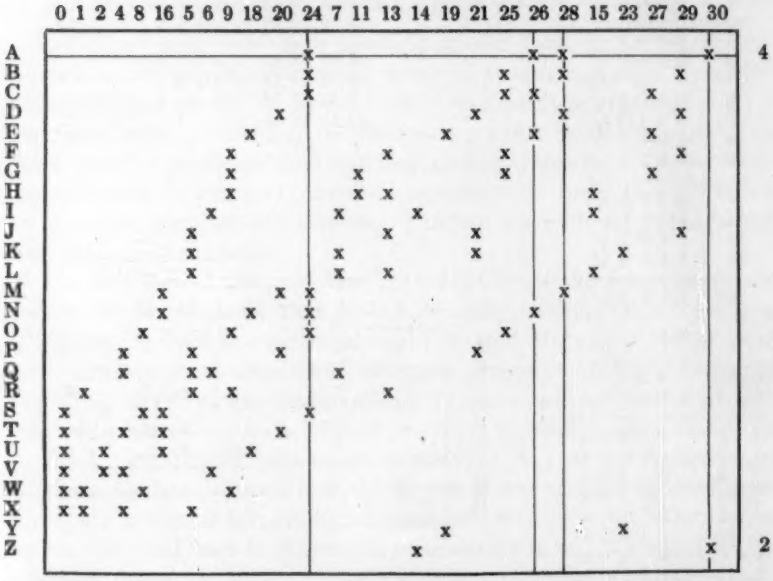
$$T_{ii} = x_4'x_3'x_1' + x_1x_2' + x_3x_2 + x_4x_2$$

of its crosses circled can be discarded without violating the requirement that each column should have at least one cross from a selected row. Rows with no circled entries are discarded (one by one, since removal of one row may require more crosses to be circled) until each selected row contains at least one circled cross. This completes the first step. The selected rows now correspond to a first approximation to a minimum sum. A check should be made to determine whether the number of selected rows is equal to the minimum number of basis rows. In Table X there are at most 4 crosses per row and 26 columns so that the minimum num-

TABLE IX — TABLE OF PRIME IMPLICANTS FOR TRANSMISSION

$T = \sum (0, 1, 2, 4, 5, 6, 7, 8, 9, 11, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30)$

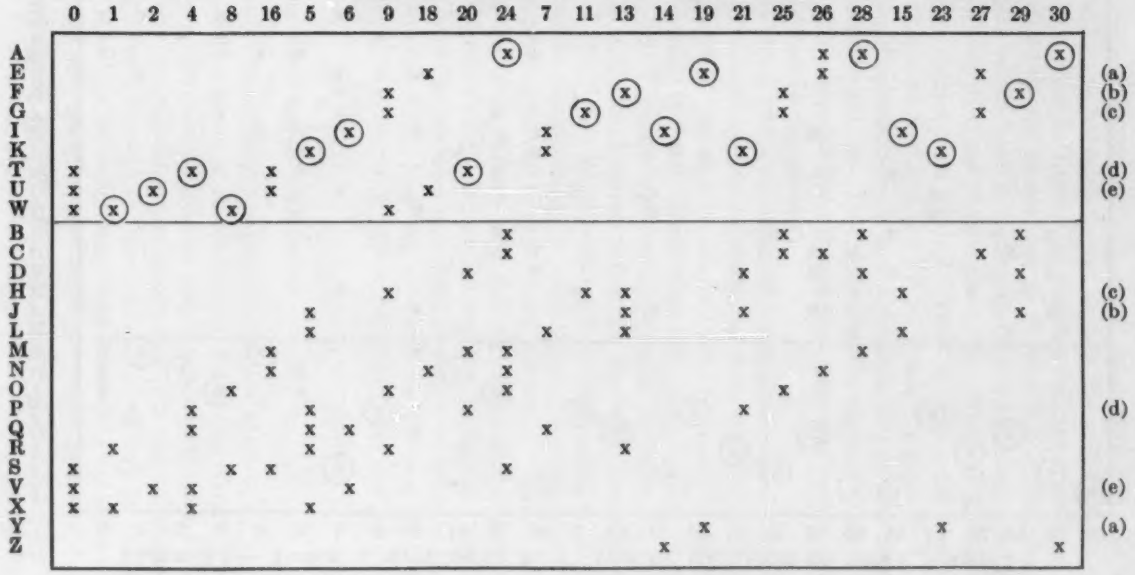
The selection of row A is shown



ber of basis rows is $[\frac{26}{4}] + 1 = 7$. Since the number of selected rows is 9 there is no guarantee that they correspond to a minimum sum.

If such an approximation to a minimum sum is not acceptable, then further work is necessary in order to reduce the number of selected rows. For each of the selected rows, a check is made of whether any of the rows in the lower part of the table (non-selected rows) have crosses in all columns in which the selected row has *circled* crosses. In Table X row E has a circled cross only in column 19; since row Y also has a cross in column 19 rows E and Y are labeled "a". Other pairs of rows which have the same relation are labeled with lower case letters, b, c, d, e in Table X. It is possible to interchange pairs of rows which are labeled with the same lower case letter without violating the requirement that each column must contain a cross from at least one selected row. If a non-selected row is labeled with two lower case letters then it *may* be possible to replace two selected rows by this one non-selected row and thereby reduce the

TABLE X — TABLE IX AFTER PARTITIONING



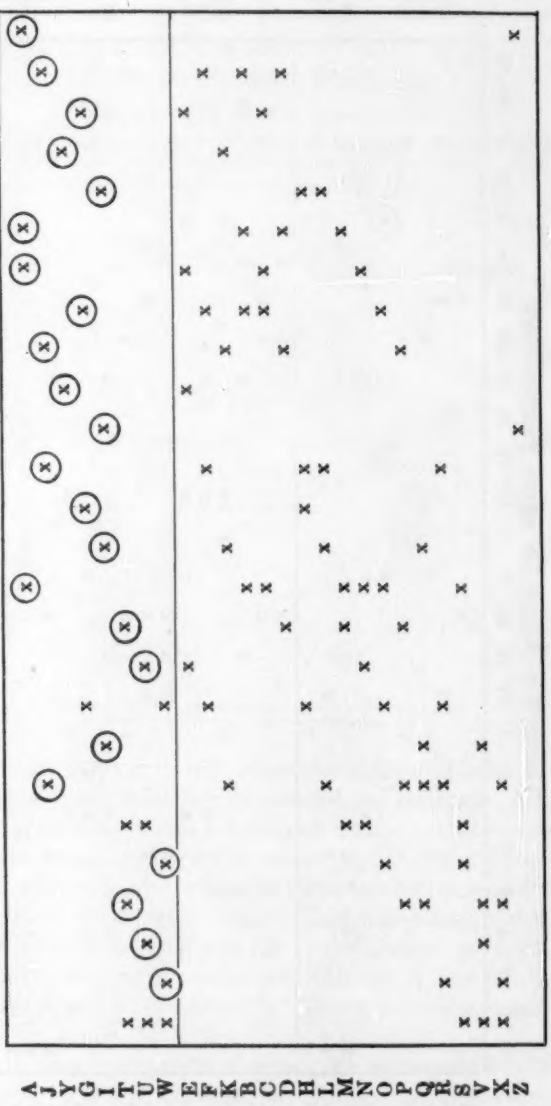


TABLE XI — TABLE X WITH ROWS E, F, AND K, REPLACED BY ROWS Y AND J
0 1 2 4 8 16 5 6 9 18 20 24 7 11 13 14 19 21 25 26 28 15 23 27 29 30
A J Y G I T U W E F K B C D H L M N O P Q R S V X Z

total number of selected rows (a check must be made that the two selected rows being removed do not contain the only two selected-row crosses in a column). In Table X no such interchange is possible.

Next a check should be made as to whether two of the labeled non-selected rows can be used to replace three selected rows, etc. In Table X rows Y(a) and J(b) can replace rows E(a), F(b) and K or rows Y(a) and P(d) can replace rows E(a), T(d) and K. The table which results from replacing rows E, F and K by rows Y and J is given in Table XI. The number of selected rows is now 8 which is still greater than 7, the minimum number possible. This table actually represents the minimum sum for this transmission even though this cannot be proved rigorously by the procedure being described.

If it is assumed that a minimum sum can always be obtained by exchanging pairs of selected and nonselected rows until it finally becomes possible to replace two or more selected rows by a single selected row, then it is possible to show directly that the Table XI does represent a minimum sum. The only interchange possible in Table XI is that of rows T and P. If this replacement is made then a table results in which only rows J and F can be interchanged. Interchanging rows J and F does not lead to the possibility of interchanging any new pairs of rows so that this process cannot be carried any further.

On the basis of experience with this method it seems that it is not necessary to consider interchanges involving more than one non-selected row. Such interchanges have only been necessary in order to obtain alternate minimum sums; however, no proof for the fact that they are never required in order to obtain a minimum sum has yet been discovered.

9 AN ALTERNATE EXACT PROCEDURE

It is possible to represent the prime implicant table in an alternative form such as that given in Table XII(b). From this form not only the minimum sums but also *all* possible sum of products forms for the transmission which correspond to consistent row sets can be obtained systematically. For concreteness, this representation will be explained in terms of Table XII. Since column 0 has crosses only in rows B and C, any consistent row set must contain either row B or row C (or both). Similarly, column 3 requires that any consistent row set must contain either row D or row E (or both). When both columns 0 and 3 are considered they require that any consistent row set must contain either row B or row C (or both) and either row D or row E (or both). This requirement can be expressed symbolically as $(B + C)(D + E)$ where

TABLE XII — DERIVATION OF THE MINIMUM SUMS FOR THE TRANSMISSION

$$T = \sum (0, 3, 4, 5, 6, 7, 8, 10, 11)$$

(a) Table of Prime Implicants

		0	3	4	5	6	7	8	10	11
$x_4'x_1$	A			x	x	x	x			
$x_4'x_2'x_1'$	B	x		x						
$x_3'x_2'x_1'$	C	x						x		
$x_4'x_2x_1$	D		x				x			
$x_3'x_2x_1$	E		x							x
$x_4x_3'x_2$	F								x	x

(b) Boolean Representation of Table

$$(B + C)(D + E)(A + B)(A)(A)(A + D)(C)(F)(E + F)$$

(c) Consistent Row Sets

(A, C, F, D), (A, C, F, E)

$$T = x_4'x_3 + x_3'x_2'x_1' + x_4x_3'x_2 + x_4'x_2x_1$$

$$T = x_4'x_3 + x_3'x_2'x_1' + x_4x_3'x_2 + x_3'x_2x_1$$

addition stands for "or" (non-exclusive) and multiplication signifies "and." This expression can be interpreted as a Boolean Algebra expression and the Boolean Algebra theorems used to simplify it. In particular it can be "multiplied out":

$$(B + C)(D + E) = BD + BE + CD + CE$$

This form is equivalent to the statement that columns 0 and 3 require that any consistent row set must contain either rows B and D, or rows B and E, or rows C and D, or rows C and E.

The complete requirements for a consistent row set can be obtained directly by providing a factor for each column of the table. Thus for Table XII the requirements for a consistent row set can be written as:

$$(B + C)(D + E)(A + B)(A)(A)(A + D)(C)(F)(E + F)$$

By using the theorems that $A \cdot (A + D) = A$ and $A \cdot A = A$, this can be simplified to $ACF(D + E)$. Thus the two consistent row sets for this table are A, C, F, D and A, C, F, E and since they both contain the same number of rows, they both represent minimum sums. This is true only because rows D and E contain the same number of crosses. In general, each row should be assigned a weight $w = n - \log_2 k$, where n is the number of variables in the transmission being considered and

TABLE XIII — DETERMINATION OF THE MINIMUM SUMS FOR THE PRIME IMPLICANT TABLE OF TABLE VII BY MEANS OF THE BOOLEAN REPRESENTATION

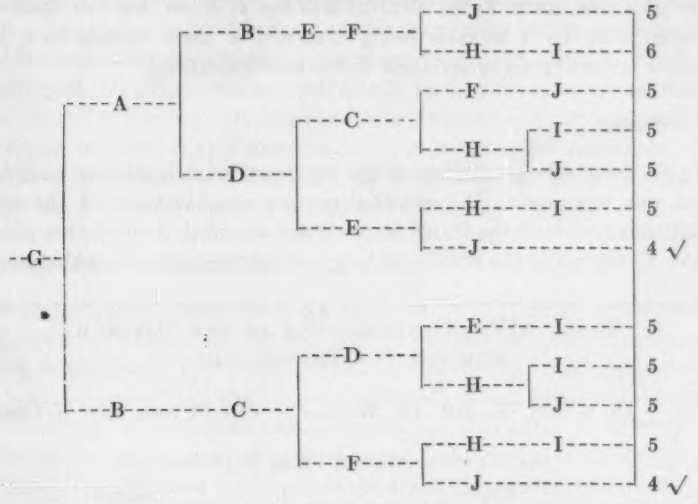
(a) Boolean representation of the Prime Implicant Table of Table VI

(A+B)(A+C)(B+D)(C+E)(D+F)(G)(E+F+H)(G+I)(H+J)(I+J)

(b) The expression of (a) after multiplying out. (The terms in italic correspond to minimum sums)

ADEJG + ACDFJG + ACDHJG + ADEHIG + ACDHIG + ABEFJG + ABEFHIG + BCDEJG + BCDHJG + BCDHIG + *BCFJG* + BCFHIG

(c) Tree circuit equivalent of (b)



k is the number of crosses in the row.* To select the minimum sums, the sum of the weights of the rows should be calculated for each row set containing the fewest rows. The row sets having the smallest total weight correspond to minimum sums. If, instead of the minimum sum, the form leading to the two-stage diode-logic circuit requiring fewest diodes is desired, a slightly different procedure is appropriate. To each row set is assigned a total weight equal to the sum of the weights of the rows plus the number of rows in the set. The desired form then corresponds to the row set having the smallest total weight.

The procedure for an arbitrary table is analogous. A more complicated example is given in Table XIII. In this example the additional

* n-$\log_2$ k is the number of literals in the prime implicant corresponding to a row containing k crosses.

theorem $(A + B)(A + C) = (A + BC)$ is useful. This example shows that for a general table the expressions described in this Section and the multiplication process can become very lengthy. However, this procedure is entirely systematic and may be suitable for mechanization.

Since the product of factors representation of a prime implicant table is a Boolean expression, it can be interpreted as the transmission of a contact network. Each consistent row set then corresponds to a path through this equivalent network. By sketching the network directly from the product of factors expression, it is possible to avoid the multiplication process. In particular the network should be sketched in the form of a tree, as in Table XIII(c) and the Boolean Algebra theorems used to simplify it as it is being drawn. For hand calculations, this method is sometimes easier than direct multiplication.

10 *d*-TERMS

In Section 1 the possibility of having *d*-entries in a table of combinations was mentioned. Whenever there are combinations of the relay conditions for which the transmission is not specified, *d*-entries are placed in the *T*-column of the corresponding rows of the table of combinations.

TABLE XIV — DETERMINATION OF THE MINIMUM SUM FOR THE TRANSMISSION

$T = \sum(5, 6, 13) + d(9, 14)$ WHERE 9 AND 14 ARE THE *d*-TERMS

(a) Determination of Prime Implicants

	$x_4x_3x_2x_1$	
5	0 1 0 1 √	
6	0 1 1 0 √	
(d) 9	1 0 0 1 √	
13	1 1 0 1 √	
(d) 14	1 1 1 0 √	

	$x_4x_3x_2x_1$
5 13	- 1 0 1
6 14	- 1 1 0
9 13	1 - 0 1

(b) Prime Implicant Table

	5	6	13
*	x		x
*		x	
			x

(c)

Basis rows: (5, 13), (6, 14)

(d)

$T = x_3x_2'x_1 + x_3x_1x_1'$

The actual values (0 or 1) of these d-entries are chosen so as to simplify the form of the transmission. This section will describe how to modify the method for obtaining a minimum sum when the table of combinations contains d-entries.

The p-terms which correspond to d-entries in the table of combinations will be called d-terms. These d-terms should be included in the list of p-terms which are used to form the prime implicants. See Table XIV. However, in forming the prime implicant table, columns corresponding to the d-terms should *not* be included, Table XIV(b). The d-terms are used in forming the prime implicants in order to obtain prime implicants containing the fewest possible literals. If columns corresponding to the d-terms were included in forming the prime implicant table this would correspond to setting all the d-entries in the table of combinations equal to 1. This does not necessarily lead to the simplest minimum sum. In the procedure just described, the d-entries will automatically be set equal to either 0 or 1 so as to produce the simplest minimum sum. For the transmission of Table XIV the 14 d-entry has been set equal to 1 and the 9 d-entry has been set equal to 0.

11 NON-CANONICAL SPECIFICATIONS

A transmission is sometimes specified not by a table of combinations or a canonical expansion, but as a sum of product terms (not necessarily prime implicants). The method described in Section 3 is applicable to such a transmission if the appropriate table of combinations (decimal specification) is first obtained. However, it is possible to modify the procedure to make use of the fact that the transmission is already partly reduced. The first step is to express the transmission in a table of binary characters such as Table XVa. Then each pair of characters is examined to determine whether any different character could have been formed from the characters used in forming the characters of the pair. For example, in Table XV(a) a (1)(0 0 0 0 1) was used in forming the (0, 1)(0 0 0 0 –) character and a (3)(0 0 0 1 1) was used in forming the (3, 7)(0 0 – 1 1) character. These can be combined to form a new character (1, 3)(0 0 0 – 1). The new characters formed by this process are listed in another column such as Table XV(b). This process is continued until no new characters are formed.

In examining a pair of characters, it is sufficient to determine whether there is only one position where one character has a one and the other character has a zero. If this is true a new character is formed which has a dash in this position and any other position in which both characters have dashes, and has a zero (one) in any position in which either charac-

TABLE XV — DETERMINATION OF THE PRIME IMPLICANTS FOR THE TRANSMISSION OF TABLE XV SPECIFIED AS A SUM OF PRODUCT TERMS

(a) Specification

		$x_5x_4x_3x_2x_1$	
0	1	0 0 0 0 –	√
0	2	0 0 0 – 0	√
3	7	0 0 – 1 1	
14	15	0 1 1 1 –	
22	23	1 0 1 1 –	
29	31	1 1 1 – 1	

(b) Characters Derived from (a)

		$x_5x_4x_3x_2x_1$	
1	3	0 0 0 – 1	√
2	3	0 0 0 1 –	√
7	15	0 – 1 1 1	√
7	23	– 0 1 1 1	√
15	31	– 1 1 1 1	√
23	31	1 – 1 1 1	√

(c) Characters Derived from (a) and (b)

	$x_5x_4x_3x_2x_1$
0 1 2 3	0 0 0 – –
7 15 23 31	– – 1 1 1

ter has a zero (one). In Table XVa the (0, 1) character has a zero in the x_2-position while the (3, 7) character has a one in the x_2-position. A new character is formed (1, 3) which has a dash in the x_2-position.

This rule for constructing new characters is actually a generalization of the rule used in Section 3 and corresponds to the theorem.

$$x_1x_2 + x_1'x_3 = x_1x_2 + x_1'x_3 + x_2x_3 .$$

Repeated application of this rule will lead to the complete set of prime implicants. As described in Section 3, any character which has all of the numbers of its decimal label appearing in the label of another character should be checked. The unchecked characters then represent the prime implicants. The process described in this section was discussed from a slightly different point of view by Quine.[7]

12 SUMMARY AND CONCLUSIONS

In this paper a method has been presented for writing any transmission as a minimum sum. This method is similar to that of Quine; however, several significant improvements have been made. The notation has been simplified by using the symbols 0, 1 and – instead of primed and unprimed variables. While it is not completely new in itself, this notation is especially appropriate for the arrangement of terms used in determining the prime implicants. Listing the terms in a column which is partitioned so as to place terms containing the same number of 1's in the same partition reduces materially the labor involved in determining the prime implicants. Such a list retains some of the advantage of the arrangement of squares in the Karnaugh Chart without requiring a geometrical representation of an n-dimensional hypercube. Since the

procedure for determining the prime implicants is completely systematic it is capable of being programmed on a digital computer. The arrangement of terms introduced here then results in a considerable saving in both time and storage space over previous methods, making it possible to solve larger problems on a given computer. It should be pointed out that this procedure can be programmed on a decimal machine by using the decimal labels instead of the binary characters introduced.

A method was presented for choosing the minimum sum terms from the list of prime implicants by means of a table of prime implicants. This is again similar to a method presented by Quine. However, Quine did not give any systematic procedure for handling cyclic prime implicant tables; that is, tables with more than one cross in each column. In this paper a procedure is given for obtaining a minimum sum from a cyclic prime implicant table. In general, this procedure requires enumeration of several possible minimum sums. If a transmission has any nontrivial group invariances it may be possible to avoid enumeration or to reduce considerably the amount of enumeration necessary. A method for doing this is given.

The process of enumeration used for selecting the terms of the minimum sum from a cyclic prime implicant table is not completely satisfactory since it can be quite lengthy. In seeking a procedure which does not require enumeration, the method involving the group invariances of a transmission was discovered. This method is an improvement over complete enumeration, but still has two shortcomings. There are transmissions which have no nontrivial group invariances but which give rise to cyclic prime implicant tables. For such transmissions it is still necessary to resort to enumeration. Other transmissions which do possess nontrivial group invariances still require enumeration after the invariances have been used to simplify the process of selecting minimum sum terms. More research is necessary to determine some procedure which will not require any enumeration for cyclic prime implicant tables. Perhaps the concept of group invariance can be generalized so as to apply to all transmissions which result in cyclic prime implicant tables.

13 ACKNOWLEDGEMENTS

The author wishes to acknowledge his indebtedness to Professor S. H. Caldwell, Professor D. A. Huffman, Professor W. K. Linvill, and S. H. Unger with whom the author had many stimulating discussions. Thanks are due also to W. J. Cadden, C. Y. Lee, and G. H. Mealy for their helpful comments on the preparation of this paper.

This research was supported in part by the Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research.

BIBLIOGRAPHY

1. Karnaugh, M., The Map Method for Synthesis of Combinational Logic Circuits, Trans. A.I.E.E., **72**, Part I pp. 593–598, 1953.
2. Keister, W., Ritchie, A. E., Washburn, S., The Design of Switching Circuits, New York, D. Van Nostrand Company, Inc., 1951.
3. Shannon, C. E., A Symbolic Analysis of Relay and Switching Circuits, Trans. A.I.E.E., **57**, pp. 713–723, 1938.
4. Shannon, C. E., The Synthesis of Two-Terminal Switching Circuits, B.S.T.J., **28**, pp. 59–98, 1949.
5. Staff of the Harvard Computation Laboratory, Synthesis of Electronic Computing and Control Circuits, Cambridge, Mass., 1951, Harvard University Press.
6. Quine, W. V., The Problem of Simplifying Truth Functions, The American Mathematical Monthly, **59**, No. 8, pp. 521–531, Oct., 1952.
7. Quine, W. V., A Way to Simplify Truth Functions, The American Mathematical Monthly, **62**, pp. 627–631, Nov., 1955.

Detection of Group Invariance or Total Symmetry of a Boolean Function*

By E. J. McCLUSKEY, Jr.

(Manuscript received June 26, 1956)

A method is presented for determining whether a Boolean function possesses any group invariance; that is, whether there are any permutations or primings of the independent variables which leave the function unchanged. This method is then extended to the detection of functions which are totally symmetric.

1 GROUP INVARIANCE

For some Boolean transmission functions (transmissions, for short) it is possible to permute the variables, or prime some of the variables, or both permute and prime variables without changing the transmission. The following material presents a method for determining, for any given transmission, which of these operations (if any) can be carried out without changing the transmission.

The permutation operations will be represented symbolically as follows:

$S_{123\ldots n}T$ will represent the transmission T with no variables permuted

$S_{213\ldots n}T$ will represent the transmission T with the x_1 and x_2 variables interchanged, etc.

Thus $S_{1432}T(x_1 , x_2 , x_3 , x_4) = T(x_1 , x_4 , x_3 , x_2)$

The symbolic notation for the priming operation will be as follows:

$N_{0000\ldots0}T$ will represent the transmission T with no variables primed

$N_{0110\ldots0}T$ will represent the transmission T with the x_2 and x_3 variables primed, etc.

Thus $N_{1010}T(x_1 , x_2 , x_3 , x_4) = T(x_1', x_2 , x_3', x_4)$.

The notation for the priming operator can be shortened by replacing the binary subscript on N by its decimal equivalent. Thus N_9T is equiv-

* This paper is derived from a thesis submitted to the Massachusetts Institute of Technology in partial fulfillment of the requirements for the degree of Doctor of Science on April 30, 1956.

TABLE I — TRANSMISSION MATRICES SHOWING EFFECT OF INTERCHANGING OR PRIMING VARIABLES

(a) Transmission Matrix

	x_1	x_2	x_3	x_4
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
9	1	0	0	1
10	1	0	1	0
11	1	0	1	1

(b) Transmission Matrix with x_3 and x_4 columns interchanged

	x_1	x_2	x_4	x_3
0	0	0	0	0
2	0	0	1	0
1	0	0	0	1
10	1	0	1	0
9	1	0	0	1
11	1	0	1	1

(c) Transmission Matrix with entries of the x_3 and x_4 columns primed

	x_1	x_2	x_3'	x_4'
3	0	0	1	1
2	0	0	1	0
1	0	0	0	1
10	1	0	1	0
9	1	0	0	1
8	1	0	0	0

alent to $N_{1001}T$. The permutation and priming operators can be combined. For example,

$$S_{2134}N_3T(x_1, x_2, x_3, x_4) = T(x_2, x_1, x_3', x_4')$$

The symbols S_iN_j form a mathematical group,[1] hence the term group invariance.

The problem considered here is that of determining which N_i and S_j satisfy the relation $N_iS_jT = T$ for a given transmission T. Since there are only a finite number of different N_i and S_j operators it is possible in principle to compute N_iS_jT for all possible N_iS_j and then select those N_iS_j for which $N_iS_jT = T$. If T is a function of n variables, there are $n!$ possible S_j operators and 2^n N_i operators so that there are $n!2^n$ possible combinations of N_iS_j. Actually, if $N_iS_jT = T$ then N_iT must equal $S_jT^{(2)}$ so that it is only necessary to compute all N_iT and all S_jT. For $n = 4$, $n! = 24$ and $2^n = 16$ so that the number of possibilities to be considered is quite large even for functions of only four variables. It is possible to avoid enumerating all N_iT and S_jT by taking into account certain characteristics of the transmission being considered.

The first step in determining the group invariances of a transmission is the same as that for finding the prime implicants.* The binary equivalents of the decimal numbers which specify the transmission are listed as in Table I(a). This list of binary numbers will be called the *transmission matrix*. When two variables are interchanged, the corresponding columns of the transmission matrix are also interchanged, Table I(b). When a variable is primed, the entries in the corresponding column of the transmission matrix are also primed, 0 replaced by 1 and 1 replaced by 0, Table I(c).

If an N_iS_j operation leaves a transmission unchanged then the cor-

* Minimization of Boolean Functions, see page 1417 of this issue.

responding matrix operations will not change the transmission matrix aside from possibly reordering the rows. In other words, it should be possible to reorder the rows of the modified transmission matrix to regain the original transmission matrix. The matrices of Table I(a) and (b) are identical except for the interchange of the 1 and 2 and the 9 and 10 rows. It is not possible to make the matrix of Table I(c) identical with that of Table I(a) by reordering rows; therefore the operation of priming the x_3 and x_4 variables does not leave the transmission $T = \sum (0, 1, 2, 9, 10, 11)$ unchanged.

If interchanging two columns of a matrix does not change the matrix aside from rearranging the rows, then the columns which were interchanged must both contain the same number of 1's (and 0's). This must

TABLE II — PARTITIONING OF THE STANDARD MATRIX FOR $T = \sum (4, 5, 7, 8, 9, 11, 30, 33, 49)$

(a) Transmission Matrix

	x_1	x_2	x_3	x_4	x_5	x_6
4	0	0	0	1	0	0
8	0	0	1	0	0	0
5	0	0	0	1	0	1
9	0	0	1	0	0	1
33	1	0	0	0	0	1
7	0	0	0	1	1	1
11	0	0	1	0	1	1
49	1	1	0	0	0	1
30	0	1	1	1	1	0
Number of 0's	7	7	5	5	6	3
Number of 1's	2	2	4	4	3	6

(b) Standard Matrix for (a) Matrix

	x_1	x_2	x_3	x_4	x_5	x_6'	Weight
4	0	0	0	1	0	0	1
8	0	0	1	0	0	0	1
32	1	0	0	0	0	0	1
5	0	0	0	1	0	1	2
6	0	0	0	1	1	0	2
9	0	0	1	0	0	1	2
10	0	0	1	0	1	0	2
48	1	1	0	0	0	0	2
31	0	1	1	1	1	1	5
	7	7	5	5	6	6	
	2	2	4	4	3	3	

(c) Second Partitioning of rows for (b) matrix

x_1	x_2	x_3	x_4	x_5	x_6'
0	0	0	1	0	0
0	0	1	0	0	0
1	0	0	0	0	0
0	0	0	1	0	1
0	0	0	1	1	0
0	0	1	0	0	1
0	0	1	0	1	0
1	1	0	0	0	0
0	1	1	1	1	1

(d) Final Partitioning for (b) matrix

x_1	x_2	x_3	x_4	x_5	x_6'
0	0	0	1	0	0
0	0	1	0	0	0
1	0	0	0	0	0
0	0	0	1	0	1
0	0	0	1	1	0
0	0	1	0	0	1
0	0	1	0	1	0
1	1	0	0	0	0
0	1	1	1	1	1

be true since rearranging the rows of a matrix does not change the total number of 1's in each column. Similarly, if priming some columns of a matrix leaves the rows unchanged, either each column must have an equal number of 1's and 0's or else for each primed column which has an unequal number of 0's and 1's there must be a second primed column which has as many 1's as the first primed column has 0's and vice versa. Such pairs of columns must also be interchanged to keep the total number of 1's in each column invariant. For the matrix of Table II(a) the only operations that need be considered are either interchanging x_1 and x_2 or x_3 and x_4 or priming and interchanging x_5 and x_6.

For the present it will be assumed that no columns of the matrix have an equal number of 0's and 1's. It is possible to determine all permuting and priming operations which leave such a matrix unchanged by considering only permutation operations on a related matrix. This related matrix, called the *standard matrix*, is formed by priming all the columns of the original matrix which have more 1's than 0's, the x_6 column in the matrix of Table II(a). Each column of a standard matrix must contain more 0's than 1's, Table II(b). The N_iS_j operations which leave the original matrix unchanged can be determined directly from the operations that leave the corresponding standard matrix unchanged. It is therefore only necessary to consider standard matrices.

Since no columns of a standard matrix have an equal number of 1's and 0's and no columns have more 1's than 0's it is only necessary to consider permuting operations. The number of 1's in a column (or row) will be called the *weight* of the column (or row). Only columns or rows which have the same weights can be interchanged. The matrix should be partitioned so that all columns (or rows) in the same partition have the same weight, Table II(b). It is now possible to interchange columns in the same column partition and check whether pairs of rows from the same row partition can then be interchanged to regain the original matrix. This can usually be done by inspection. For example, in Table II(b) if columns x_4 and x_3 are interchanged, then interchanging rows 4 and 8, 5 and 9, and 6 and 10 will regain the original matrix.

The process of inspection can be simplified by carrying the partitioning further. In the matrix of Table II(b), row 32 cannot be interchanged with either row 8 or row 4. This is because it is not possible to make row 32 identical with either row 8 or row 4 by interchanging columns x_1 and x_2. Row 32 has weight 1 in these columns while rows 8 and 14 both have weight 0. In general, only rows which have the same weight in *each* submatrix can be interchanged. Permuting columns of the same partition does not change the weight of the rows in the corresponding submatrices.

The matrix can therefore be further partitioned by separating the rows into groups of rows which have the same weight in every column partition, Table II(c). Similar remarks hold for the columns so that it may then be necessary to partition the columns again so that each column in a partition has the same weight in each submatrix, Table II(d). Partitioning the columns may make it necessary to again partition the rows, which in turn may make more column partitioning necessary. This process should be carried out until a matrix results in which each row (column) of each submatrix has the same weight. Inspection is then used to determine which row and column permutations will leave the matrix unchanged. Only permutations among rows or columns in the same partition need be considered.

From the matrix of Table II(d) it can be seen that permuting either columns x_3 and x_4 or columns x_5 and x_6' will not change the matrix aside from reordering certain rows. This means that interchanging x_3 and x_4 or priming and interchanging x_5 and x_6 in the original transmission will leave the transmission unchanged. Interchanging x_6' and x_5 means replacing x_5 by x_6' and x_6 by x_5' which is the same as interchanging x_5 and x_6 and then priming both x_5 and x_6. Thus for the transmission of Table II $S_{124356}T = T$ and $N_{000011}S_{123465}T = N_3S_{123465}T = T$.

A procedure has been presented for determining the group invariance of any transmission matrix which does not have an equal number of 1's and 0's in any column. This must now be extended to matrices which do have equal numbers of 0's and 1's in some columns, Table III(a). For such matrices the procedure is to prime appropriate columns so that there are either more 0's than 1's or the same number of 0's and 1's in each column, Table III(a). This matrix is then partitioned as described above and the permutations which leave the matrix unchanged are determined. The matrix of Table III(a) is so partitioned. Interchanging

TABLE III — TRANSMISSION MATRICES FOR $T = \sum (0, 6, 9, 12)$

(a) Transmission Matrix

	x_1	x_2	x_3	x_4
0	0	0	0	0
6	0	1	1	0
9	1	0	0	1
12	1	1	0	0
Number of 0's	2	2	3	3
Number of 1's	2	2	1	1

(b) Transmission Matrix with x_1 and x_2 primed

	x_1'	x_2'	x_3	x_4
0	0	0	0	0
10	1	0	1	0
5	0	1	0	1
12	1	1	0	0
	2	2	3	3
	2	2	1	1

both x_1 and x_2, and x_3 and x_4 leave this matrix unchanged so that $S_{2143}T = T$. The possibility of priming different combinations of the columns which have an equal number of 0's and 1's must now be considered. Certain of the possible combinations can be excluded beforehand. In Table III(a) the only possibility which must be considered is that of priming both x_1 and x_2. If only x_1 or x_2 is primed, there will be no row which has all zeros. No permutation of the columns of this matrix (with x_1 or x_2 primed) can produce a row with all zeros. Therefore this matrix cannot possibly be made equal to the original matrix by rearranging rows and columns. Priming both x_1 and x_2 must be considered since the 12-row will be converted into a row with all zeros. The operation of priming x_1 and x_2 is written symbolically as $N_{1100} = N_{12}$. In general, if the matrix has a row consisting of all zeros, only those N_i operations for which i is the number of some row in the matrix, need be considered. If the row does not have an all-zero row, only those N_i for which i is *not* the number of some row need be considered. Similarly, if the matrix has a row consisting of all 1's, only those N_i for which there is some row of the matrix which will be converted into an all-one row, need be considered. This is equivalent to considering only those N_i for which some row has a number $k = 2^n - 1 - i$* where n is the number of columns. If the matrix does *not* have an all-one row, only those N_i for which *no* row has a number $k = 2^n - 1 - i$ should be considered.

Each priming operation which is not excluded by these rules is applied to the transmission matrix. The matrices so formed are then partitioned as described previously. Any of these matrices that have the same partitioning as the original matrix are then inspected to see if any row and column permutations will convert them to the original matrix. For the matrix of Table III(a) the operation of priming both x_1 and x_2 was not excluded. The matrix which results when these columns are primed is shown in Table III(b). Inspection of this figure shows that interchange of either x_3 and x_4 or x_1' and x_2' will convert the matrix back to the matrix of Table III(a). Therefore, for the transmission of this table $S_{1243}N_{1100}T = T$ and $S_{2134}N_{1100}T = T$.

2 TOTAL SYMMETRY

There are certain transmissions whose value depends not on which relays are operated but only on how many relays are operated. For

* The number of the row which has all ones is $2^n - 1$. If N_i operating on some row, k, is to produce the all-one row, i must have 1's wherever k has 0's and vice versa. This means that

$$i + k = 2^n - 1 \quad \text{or} \quad k = 2^n - 1 - i.$$

TABLE IV — TRANSMISSION MATRIX FOR
$T = \sum (3, 5, 6, 7) = S_{2,3}(x_1, x_2, x_3)$

	x_3	x_2	x_1
3	0	1	1
5	1	0	1
6	1	1	0
7	1	1	1

example, the transmission of Table IV equals 1 whenever two or three relays are operated. For such transmissions any permutation of the variables leaves the transmission unchanged. These transmissions are called *totally symmetric.*[3] They are usually written in the form, $T = S_{a_1, a_2 \cdots a_m}(x_1, x_2, \cdots x_n)$, where the transmission is to equal 1 only when exactly a_1 or a_2 or $\cdots$ or a_m of the variables $x_1, x_2 \cdots x_n$ are equal to one. The transmission of Table IV can be written as $S_{2,3}(x_1, x_2, x_3)$. This definition of symmetric transmissions can be generalized by allowing some of the variables $(x_1, x_2, \cdots x_n)$ to be primed. Thus the transmission $S_2(x_1, x_2', x_3)$ will equal 1 only when $x_1 = x_2' = x_3 = 1$ or $x_1 = x_3 = 1$ and $x_2 = 0$. It is useful to know when a transmission is totally symmetric since special design techniques exist for such functions.[4]

It is possible to determine whether a transmission is totally symmetric from its matrix. Unless all columns of the standard matrix derived from the transmission matrix have the same weight, the transmission cannot possibly be totally symmetric. If all columns do have equal weights, the rows should be partitioned into groups of rows which all have the same weight. Whether the transmission is totally symmetric can now be determined by inspection. If there is a row of weight k; that is, a row which contains k 1's, then every possible row of weight k must also be included in the matrix. This means that there must be ${}_nC_k$ rows of weight k where n is the number of columns (variables).* If any possible row of weight k was not included then the corresponding k literals could be set equal to 1 without the transmission being equal to 1. This contradicts the definition of a totally symmetric transmission. In Table V(b) there are 4 rows of weight 1 and 1 row of weight 4. Since ${}_4C_1 = 4$ and ${}_4C_4 = 1$ this transmission is totally symmetric and can be written as $S_{1,4}(x_1, x_2', x_3, x_4')$. The number of rows of weight 1 in Table V(d) is 2 and since ${}_4C_1 = 4$ this transmission is *not* totally symmetric.

A difficulty arises if all columns of a transmission matrix contain equal

* ${}_nC_k$ is the binomial coefficient $\frac{n!}{(n-k)!k!}$

TABLE V — DETERMINATION OF TOTALLY SYMMETRIC TRANSMISSION

(a) Transmission Matrix for $T = \sum (1, 4, 7, 10, 13)$

	x_1	x_2	x_3	x_4
1	0	0	0	1
4	0	1	0	0
10	1	0	1	0
7	0	1	1	1
13	1	1	0	1
Number of 0's	3	2	3	2
Number of 1's	2	3	2	3

(b) Standard Matrix for $T = \sum (1, 4, 7, 10, 13)$ showing that $T = S_{1,4}(x_1, x_2', x_3, x_4')$

	x_1	x_2'	x_3	x_4'
1	0	0	0	1
2	0	0	1	0
4	0	1	0	0
8	1	0	0	0
15	1	1	1	1
	3	3	3	3
	2	2	2	2

(c) Transmission Matrix for $T = \sum (3, 5, 10, 12, 13)$

	x_1	x_2	x_3	x_4
3	0	0	1	1
5	0	1	0	1
10	1	0	1	0
12	1	1	0	0
13	1	1	0	1
Number of 0's	2	2	3	2
Number of 1's	3	3	2	3

(d) Standard Matrix for $T = \sum (3, 5, 10, 12, 13)$ showing that it is not totally symmetric

	x_1'	x_2'	x_3	x_4'
0	0	0	0	0
1	0	0	0	1
8	1	0	0	0
7	0	1	1	1
14	1	1	1	0
	3	3	3	3
	2	2	2	2

TABLE VI — DETERMINATION OF TOTAL SYMMETRY FOR $T = \sum (0, 3, 5, 10, 12, 15)$

(a) Transmission Matrix for $T(x_1, x_2, x_3, x_4)$

	x_1	x_2	x_3	x_4
0	0	0	0	0
3	0	0	1	1
5	0	1	0	1
10	1	0	1	0
12	1	1	0	0
15	1	1	1	1
Number of 0's	3	3	3	3
Number of 1's	3	3	3	3

(b) Standard Matrix for $T(1, x_2, x_3, x_4)$

	x_2'	x_3'	x_4
	1	0	0
	0	1	0
	0	0	1
Number of 0's	2	2	2
Number of 1's	1	1	1

$T(1, x_2, x_3, x_4) = S_1(x_2', x_3', x_4)$

(c) Standard Matrix for $T(0, x_2, x_3, x_4)$

	x_2	x_3	x_4'
	0	0	1
	0	1	0
	1	0	0
Number of 0's	2	2	2
Number of 1's	1	1	1

$T(0, x_2, x_3, x_4) = S_1(x_2, x_3, x_4') = S_2(x_2', x_3', x_4)$

numbers of zeros and ones as in Table VI(a). For such a matrix it is not clear which variables should be primed. It is possible to avoid considering all possible primings by "expanding" the transmission about one of the variables by means of the theorem

$$T(x_1, x_2, \cdots x_n) = x_1 T(1, x_2, \cdots x_n) + x_1' T(0, x_2, \cdots x_n)^{2,3}$$

and then making use of the relation:

$$S_{a_1, a_2, \cdots a_m}(x_1, x_2, \cdots x_n) = x_1 S_{a_1-1, a_2-1, a_3-1, \cdots a_m-1}(x_2, \cdots x_m) + x_1' S_{a_1, a_2, \cdots a_m}(x_2, \cdots x_n)^5$$

This technique is illustrated in Table VI. The standard matrix for $T(1, x_2, x_3, x_4)$ has three rows each containing a single one so that

$$T(1, x_2, x_3, x_4) = S_1(x_2', x_3', x_4)$$

The transmission $T(0, x_2, x_3, x_4)$ has an identical standard matrix so that

$$T(0, x_2, x_3, x_4) = S_1(x_2, x_3, x_4')$$

This can be written in terms of x_2', x_3', and x_4 :

$$S_1(x_2, x_3, x_4') = S_2(x_2', x_3', x_4)^5.$$

Finally

$$T(x_1, x_2, x_3, x_4) = x_1 T(1, x_2, x_3, x_4) + x_1' T(0, x_2, x_3, x_4)$$
$$= x_1 S_1(x_2', x_3', x_4) + x_1' S_2(x_2', x_3', x_4) = S_2(x_1, x_2', x_3', x_4).*$$

The method just presented for detecting total symmetry is more systematic than the only other available method[5] and applies for transmissions of any number of variables.

BIBLIOGRAPHY

1. Birkhoff, G., and MacLane, S., A Survey of Modern Algebra, The MacMillan Company, New York.
2. Shannon, C. E., The Synthesis of Two-Terminal Switching Circuits, B.S.T.J., **28**, pp. 59–98, 1949.
3. Shannon, C. E., A Symbolic Analysis of Relay and Switching Circuits, Trans. A.I.E.E., **57**, pp. 713–723, 1938.
4. Keister, W., Ritchie, A. E., Washburn, S., The Design of Switching Circuits, New York, D. Van Nostrand Company, Inc., 1951.
5. Caldwell, S. H., The Recognition and Identification of Symmetric Switching Circuits, Trans. A.I.E.E., **73**, Part I, pp. 142–146, 1954.

* This technique for transmission matricies having an equal number of zeros and ones in all columns was brought to the author's attention by Wayne Kellner, a student at the Massachusetts Institute of Technology.

Bell System Technical Papers Not Published in This Journal

ANDERSON, O. L.[1]

Effect of Pressure on Glass Structure, J. Appl. Phys., **27,** pp. 943–949, Aug., 1956.

ANDERSON, P. W.[1]

Ordering and Antiferromagnetism in Ferrites, Phys. Rev., **102,** pp. 1008–1013, May 15, 1956.

ANDERSON, P. W., see Holden, A. N.

ARNOLD, S. M.,[1] and KOONCE, S. ELOISE[1]

Filamentary Growths of Metals at Elevated Temperatures, J. Appl. Phys., Letter to the Editor, **27,** p. 964, Aug., 1956.

BONNEVILLE, S., See Noyes, J. W.

BRIDGERS, H. E.[1]

The Formation of P–N Junctions in Semiconductors by the Variation of Crystal Growth Parameters, J. Appl. Phys., **27,** pp. 746–751, July, 1956.

BOZORTH, R. M.,[1] WILLIAMS, H. J.,[1] and WALSH, DOROTHY E.[1]

Magnetic Properties of Some Orthoferrites and Cyanides at Low Temperatures, Phys. Rev., **103,** pp. 572–578, August 1, 1956.

CHASE, F. H.[1]

Power Regulation by Semiconductors, Elec. Engg., **75,** pp. 818–822, Sept., 1956.

CHEN, W. H., see Lee, C. Y.

---

[1] Bell Telephone Laboratories, Inc.

CHYNOWETH, A. G.[1]

Spontaneous Polarization of Guanidine Aluminum Sulfate Hexahydrate at Low Temperatures, Phys. Rev., **102,** pp. 1021–1023, May 15, 1956.

COOK, R. K.[1] and WASILIK, J. H.[1]

Anelasticity and Dielectric Loss of Quartz, J. Appl. Phys., **27,** pp. 836–837, July, 1956.

DARROW, K. K.[1]

Electron Physics in America, Physics Today, **9,** pp. 23–27, Aug., 1956

DAVID, E. E., JR.[1]

Naturalness and Distortion in Speech Processing Devices, J. Acous. Soc. Am., **28,** pp. 586–589, July, 1956.

DAVID, E. E., JR.,[1] and MCDONALD, H. S.[1]

A Bit-Squeezing Technique Applied to Speech Signals, I.R.E. Convention Record, **4, Part 4,** pp. 148–153, July, 1956.

DEWALD, J. F.[1] and LEPOUTRE, G.[1]

I — The Thermoelectric Properties of Metal-Ammonia. II — The Thermoelectric Power of Sodium and Potassium Solutions at −78° and the Effect of Added Salt on the Thermoelectric Power of Sodium at −33°. III — Theory and Interpretation of Results, J. Am. Chem. Soc., **78,** pp. 2953–2962, July 5, 1956.

EDER, M. J., see Veloric, H. S.

EMBREE, M. L.,[1] and WILLIAMS, D. E.[1]

An Automatic Card Punching Transistor Test Set, Proc. 1956 Electronic Components Symposium, pp. 125–130, 1956.

FARRAR, H. K., see Maxwell, J. L.

FEHER, G.[1]

Method of Polarizing Nuclei in Paramagnetic Substances, Phys. Rev., Letter to the Editor, **103,** pp. 500–501, July 15, 1956.

---

[1] Bell Telephone Laboratories, Inc.

FEHER, G.,[1] and GERE, E.[1]

Polarization of Phosphorus Nuclei in Silicon, Phys. Rev., Letter to the Editor, **103,** pp. 501–503, July 15, 1956.

FREYNIK, H. S., see Gohn, G. R.

FTHENAKIS, E.[1]

A Voltage Regulator Using High Speed of Response Magnetic Amplifiers With Transistor Driver, Proc. Special Tech. Conf. on Magnetic Amplifiers, **T-86,** pp. 185–199, July, 1955.

GAUDET, G., see Noyes, J. W.

GELLER, S.,[1] and WOOD, Mrs. E. A.,[1]

Crystallographic Studies of Perovskite-Like Compounds. I — Rare Earth Orthoferrites and $YFeO_3$, $YCrO_3$, $YAlO_3$, Acta Crys., **9,** pp. 563–568, July 10, 1956.

GERE, E., see Feher, G.

GIANOLA, U. F.,[1] and JAMES, D. B.[1]

Ferromagnetic Coupling Between Crossed Coils, J. Appl. Phys., **27,** pp. 608–609, June, 1956.

GILBERT, E. N.[1]

Enumeration of Labelled Graphs, Canadian J. of Math., **8,** pp. 405–411, 1956.

GOHN, G. R.[1]

Fatigue and Its Relation to the Mechanical and Metallurgical Properties of Metals, SAE Trans., **64,** pp. 31–40, 1956.

GOHN, G. R.,[1] FREYNIK, H S.,[2] and GUERARD, J. P.[1]

The Mechanical Properties of Wrought Phosphor Bronze Alloys, A.S.T.M. Special Tech. Pub., **STP 183,** pp. 1–114, Jan., 1956.

GUERARD, J. P., see Gohn G. R.

HANNAY, N. B.[1]

Recent Advances in Silicon — Progress in Semiconductors, Book, 1, pp. 1–35, 1956. (Published by Heywood & Co., Ltd., London)

[1] Bell Telephone Laboratories, Inc.
[2] Riverside Metal Co., Div., H. K. Porter Co., Inc., Riverside, N. J.

HOLDEN, A. N.,[1] MATTHIAS, B. T.,[1] ANDERSON, P. W.,[1] and LEWIS, H. W.[1]

New Low-Temperature Ferromagnets, Phys. Rev., **102,** p. 1463, June 15, 1956.

HUNTLEY, H. R.[2]

The Present and Future of Telephone Transmission, Elec. Engg., **75,** pp. 686–692, Aug., 1956.

JAMES, D. B., see Gianola, U. F.

JONES, H. L.[9]

A Blend of Operations Research and Quality Control in Balancing Loads on Telephone Equipment, Trans. Am. Soc. Quality Control (1956 Montreal Convention).

KAMINOW, I. P., see Kircher, R. J.

KIRCHER, R. J.[1] and KAMINOW, I. P.[1]

Super-Regenerative Transistor Oscillator, Electronics, **29,** pp. 166–167, July, 1956.

KRETZMER, E. R.[1]

Reduced-Alphabet Representation of TV Signals, I.R.E. Convention Record, **4, Part 4,** pp. 140–147, 1956.

KOONCE, S. ELOISE, see Arnold, S. M.

LEE, C. Y.,[1] and CHEN, W. H.[4]

Several-Valued Combinational Switching Circuits, Commun. and Electronics, **25,** pp. 278–283, July, 1956.

LEPOUTRE, G., see Dewald, J. F.

LEWIS, H. W.[1]

Two-Fluid Model of an "Energy-Gap" Superconductor, Phys. Rev., **102,** pp. 1508–1511, June 15, 1956.

LEWIS, P. W., see Holden, A. N.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[4] University of Florida, Gainesville, Fla.
[9] Illinois Bell Telephone Company, Chicago, Ill.

MANLEY, J. M.,[1] and ROWE, H. E.[1]

Some General Properties of Non-Linear Elements. Part 1 — General Energy Relations, Proc. I.R.E., **44,** pp. 904–913, July, 1956.

MATTHIAS, B. T., see Holden, A. N.; Wood, E. A.

MAXWELL, J. L.,[6] and FARRAR, H. K.[6]

Automatic Dispatch System for Teletypewriter Lines, Elec. Engg., **75,** p. 705, Aug., 1956.

McDONALD, H. S., see David E. E.

McCLEAN, D. A.[1] and POWER, F. S.[1]

Tantalum Solid Electrolytic Capacitors, Proc. I.R.E., **44,** pp. 872–878, July, 1956.

McMAHON, W.[1]

Dielectric Effects Produced by Solidifying Certain Organic Compounds in Electric or Magnetic Fields, J. Am. Chem. Soc., **78,** pp. 3290–3294, July 20, 1956.

MERZ, W. J.[1]

Effect of Hydrostatic Pressure on the Hysteresis Loop of Guanidine Aluminum Sulfate Hexahydrate, Phys. Rev., **103,** pp. 565–566, Aug. 1, 1956.

MERZ, W. J.[1]

Switching Time in Ferroelectric $BaTiO_3$ and Its Dependence on Crystal Thickness, J. Appl. Phys., **27,** pp. 938–943, Aug. 1, 1956.

NELSON, L. S.[1]

Windowed Dewar Vessels for Use at Low Temperatures, Rev. Sci. Instr., **27,** pp. 655–656, Aug., 1956.

NOYES, J. W.,[5] GAUDET, G.,[5] and BONNEVILLE, S.[5]

Development of Transcontinental Communications in Canada, Commun. and Electronics, **25,** pp. 342–352, July, 1956.

---

[1] Bell Telephone Laboratories, Inc.
[5] Bell Telephone Company of Canada, Ltd., Montreal, Que., Canada.
[6] Pacific Telephone and Telegraph Co., San Francisco, Calif.

PILLIOD, J. J.[2]

Clinton R. Hanna 1955 Lamme Medalist — History of the Metal, Elec. Engg., **75,** p. 706, Aug., 1956.

POWER, F. S., see McLean, D. A.

PRINCE, M. B., see Veloric, H. S.

RINEY, T. D.[1]

On the Coefficients in Asymptotic Factorial Expansions, Proc. of Am. Math. Soc., **7,** pp. 245–249, Apr., 1956.

ROWE, H. E., see Manley, J. M.

SHULMAN, R. G.[1]

Hole Trapping in Germanium Bombarded by High-Energy Electrons, Phys. Rev., **102,** pp. 1451–1455, June 15, 1956.

SHULMAN, R. G.,[1] and WYLUDA, B. J.[1]

Copper in Germanium; Recombination Center and Trapping Center, Phys. Rev., **102,** pp. 1455–1457, June 15, 1956.

SLICHTER, W. P.[1]

On the Morphology of Highly Crystalline Polyethylenes, J. Poly. Sci., **21,** pp. 141–143, July, 1956.

TIEN, P. K.[1]

A Dip in the Minimum Noise Figure of Beam-Type Microwave Amplifiers, Proc. I.R.E., Correspondence Sec., **44,** p. 938, July, 1956.

VELORIC, H. S.,[1] EDER, M. J.,[1] and PRINCE, M. B.[1]

Avalanche Breakdown in Silicon Diffused P-N Junctions as a Function of Impurity Gradient, J. Appl. Phys., **27,** pp. 895–899, August, 1956.

WALSH, DOROTHY E., see Bozorth, R. M.

WASILIK, J. H., see Cook, R. K.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.

WERNICK, J. H.[1]

Determination of Diffusivities in Liquid Metals by Means of Temperature-Gradient Zone-Melting, J. Chem. Phys., **25,** pp. 47–49, July, 1956.

WILKINSON, R. I.[1]

Beginnings of Switching Theory in the United States, Elec. Engg., **75,** pp. 796–802, Sept., 1956.

WILLIAMS, D. E., see Embree, M. L.

WILLIAMS, H. J., see Bozorth, R. M.

WOOD, MRS. E. A.[1]

Guanidinium Aluminum Sulfate Hexahydrate; Crystallographic Data, Acta Crys., **9,** pp. 618–619, July 10, 1956.

WOOD, MRS. E. A.[1]

The Question of a Phase Transition in Silicon, J. Phys. Chem., **60,** p. 508, 1956.

WOOD, MRS. E. A.,[1] and MATTHIAS, B. T.[1]

Crystal Structures of Nb_3Au and V_3Au, Acta Crys., **9,** pp. 534, June 10, 1956.

WOOD, E. A., see Geller, S.

WYLUDA, B. J., see Shulman, R. G.

[1] Bell Telephone Laboratories Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal

ALBRECHT, E. G., see Bullard, W. R.

ANDERSON, P. W.

Ordering and Antiferromagnetism in Ferrites, Monograph 2636.

BAKER, W. O., see Winslow, F. H.

BENNETT, W. R., see Pierce, J. R.

BOGERT, B. P.

The Vobanc — A Two-to-One Speech Bandwidth Reduction System, Monograph 2643.

BÖMMEL, H. E., MASON, W. P., and WARNER, A. W.

Dislocations, Relaxations, and Anelasticity of Crystal Quartz, Monograph 2618.

BOYET, H., see Weisbaum, S.

BULLARD, W. R., WEPPLER, H. E., ALBRECHT, E. G., DIETZ, A. E., CHRISTOFERSON, E. W., SLOTHOWER, J. E., ELLIS, H. M., PHELPS, J. W., ROACH, C. L., and TREEN, R. E.

Co-Ordinated Protection for Open-Wire Joint Use — Trends and Tests, Monograph 2662.

CHRISTOFERSON, E. W., see Bullard, W. R.

CHYNOWETH, A. G.

Spontaneous Polarization of Guanidine Aluminum Sulfate Hexahydrate at Low Temperatures, Monograph 2645.

---

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

CHYNOWETH, A. G.

Surface Space-Charge Layers in Barium Titanate, Monograph 2628.

CHYNOWETH, A. G., and MCKAY, K. G.

Photon Emission from Avalanche Breakdown in Silicon, Monograph 2619.

DACEY, G. C., see Thomas, D. E.

DANIELSON, W. E., ROSENFELD, J. L., and SALOOM, J. A.

Analysis of Beam Formation with Electron Guns of the Pierce Type, Monograph 2609.

DARLINGTON, S.

A Survey of Network Realization Techniques, Monograph 2620.

DIETZ, A. E., see Bullard, W. R.

DITZENBERGER, J. A., see Fuller, C. S.

DUDLEY, H. W.

Fundamentals of Speech Synthesis, Monograph 2648.

ELLIS, H. M., see Bullard, W. R.

FULLER, C. S., and DITZENBERGER, J. A.

Diffusion of Donor and Acceptor Elements in Silicon, Monograph 2651.

GIANOLA, U. F., and JAMES, D. B.

Ferromagnetic Coupling Between Crossed Coils, Monograph 2653.

HARROWER, G. A.

Auger Electrons in Energy Spectra of Secondary Electrons from Mo and W, Monograph 2621.

HEIDENREICH, R. D.

Thermionic Emission Microscopy of Metals, Monograph 2445.

HOLDEN, A. N., MERZ, W. J., REMEIKA, J. P., and MATTHIAS, B. T.

Properties of Guanidine Aluminum Sulfate Hexahydrate and Some of Its Isomorphs, Monograph 2580.

PRINCE, E.

Neutron Diffraction Observation of Heat Treatment in Cobalt Ferrite, Monograph 2632.

REISS, H.

P-N Junction Theory by the Method of δ-Functions, Monograph 2638

REMEIKA, J. P., see Holden, A. N.

RICE, S. O.

A First Look at Random Noise, Monograph 2659.

ROACH, C. L., see Bullard, W. R.

ROSENFELD, J. L., see Danielson, W. E.

SALOOM, J. A., see Danielson, W. E.

SLOTHOWER, J. E., see Bullard, W. R.

THEUERER, H. C.

Purification of Germanium Tetrachloride by Extraction with Hydrochloric Acid and Chlorine, Monograph 2639.

THOMAS, D. E., and DACEY, G. C.

Application Aspects of Germanium Diffused Base Transistor, Monograph 2660.

THURMOND, C. D., and LOGAN, R. A.

Copper Distribution Between Germanium and Ternary Melts Saturated with Germanium, Monograph 2640.

TREEN, R. E., See Bullard, W. R.

WARNER, A. W., see Bömmel, H. E.

WEISBAUM, S., and BOYET, H.

Broadband Nonreciprocal Phase Shifts — Two Ferrite Slabs in Rectangular Guide, Monograph 2642.

WEPPLER, H. E., see Bullard, W. R.

WINSLOW, F. H., BAKER, W. O., and YAGER W. A.

The Structure and Properties of Some Pyrolyzed Polymers, Monograph 2572.

YAGER, W. A., see Winslow, F. H.

Contributors to This Issue

C. F. Edwards, B.A. 1929 and M.A. 1930, Ohio State University; A. T. & T. Co. 1930–34; Bell Telephone Laboratories, 1935–. Research in transoceanic short wave transmission, transoceanic short wave transmission using multiple unit steerable antenna receiving system, waveguide circuit design, frequency converters for microwave radio relay systems and time division multiplex telephone system. Author of articles published in I.R.E. Proceedings. Member of I.R.E.

Joseph P. Laico, M.E., Brooklyn Polytechnic Institute, 1933; General Drafting Company, 1920–23; American Machine and Foundry Company, 1923–29; Bell Telephone Laboratories, 1929–. Supervision in the field of mechanical design and development of electronic devices is Mr. Laico's occupation at the Laboratories. He holds some twenty patents, all in electronic devices, and is a member of Tau Beta Pi.

E. J. McCluskey, Jr., A.B., 1953, Bowdoin College, B.S. and M.S. 1953 and Sc.D. 1956, M.I.T.; Bell Telephone Laboratories, co-operative student, 1950–52; M.I.T. research assistant and instructor, 1953–55; Bell Telephone Laboratories, 1955–. Research in connection with electronic switching systems. Non-resident instructor at M.I.T., summer 1956. Lecturer at C.C.N.Y., 1956. Member of I.R.E., Phi Beta Kappa, Tau Beta Pi, Eta Kappa Nu and Sigma Xi.

Hunter L. McDowell, B.E.E., Cornell University, 1948; Bell Telephone Laboratories, 1948–. At the Laboratories, Mr. McDowell has been principally engaged in vacuum tube development, particularly traveling wave amplifiers. He is a member of I.R.E.

Samuel P. Morgan, B.S. 1943, M.S. 1944 and Ph.D. 1947, California Institute of Technology; Bell Telephone Laboratories, 1947–. A research mathematician, Dr. Morgan specializes in electromagnetic theory. Studies in problems of waveguide and coaxial cable transmission and microwave antenna theory. Member of the American Physical Society, Tau Beta Pi, Sigma Xi and I.R.E.

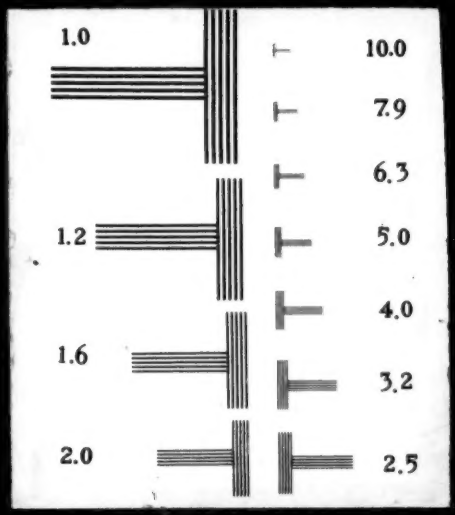
CLARENCE R. MOSTER, B.E.E., Alabama Polytechnic Institute, 1942; S.M., Massachusetts Institute of Technology, 1947; Naval Research Laboratory, 1942–45; Bell Telephone Laboratories, 1947–. Mr. Moster's main work at the Laboratories has been in vacuum tube development, specializing in microwave tubes. Member of Institute of Radio Engineers, Sigma Xi, Eta Kappa Nu and Phi Kappa Phi.

W. T. READ, JR., B.S. 1944, Rutgers and M.S. 1948, Brown University; National Defense Research Committee, 1943–46; Engaged in air-blast and earth-shock tests at Princeton University Station and measurements of air blast at Bikini atom bomb tests; Bell Telephone Laboratories, 1947–. Photoelastic and mathematical stress analysis. Dislocation theory and problems of plastic deformation were early studies. Later involved with theory of flow and space charge of holes and electrons and with electrical and mechanical effects of dislocations and other imperfections in semiconductors. Author of "Dislocations in Crystals," McGraw-Hill, 1953. Member of Phi Beta Kappa.

WILLIAM MERLIN SHARPLESS, B.S. in E.E. 1928 and Professional Engineering in E.E. 1951, University of Minnesota; Bell Telephone Laboratories, 1928–. Studies of optical behaviors of the ground for short radio waves, artificial ground systems for short wave reception, angle of arrival of transatlantic short wave signals, multiple unit steerable antenna system, microwave radio circuits, noise factors in microwave silicon rectifiers, broad band balanced and unbalanced crystal converters, radar, propagation of microwaves over land paths, angle of arrival of microwaves, and antenna systems and artificial dielectrics for microwaves. Several patents. Published papers on short radio waves and microwaves. Member of American Physical Society and Scientific Research Society of America. Senior member of I.R.E.

JAMES A. YOUNG, JR., B.S. 1943, California Institute of Technology; Radio Officer, U. S. Army Signal Corps, 1943–1946; Jet Propulsion Laboratory of California Institute of Technology, 1946–1947; Ph.D. 1953, University of Washington; Bell Telephone Laboratories, 1953–. Concerned primarily with low loss circular electric mode waveguide. Member of American Physical Society, Sigma Xi and I.R.E.

RESOLUTION CHART



100 MILLIMETERS

INSTRUCTIONS Resolution is expressed in terms of the lines per millimeter recorded by a particular film under specified conditions. Numerals in chart indicate the number of lines per millimeter in adjacent "T-shaped" groupings.

In microfilming, it is necessary to determine the reduction ratio and multiply the number of lines in the chart by this value to find the number of lines recorded by the film. As an aid in determining the reduction ratio, the line above is 100 millimeters in length. Measuring this line in the film image and dividing the length into 100 gives the reduction ratio. Example: the line is 20 mm. long in the film image, and $100/20 = 5$.

Examine "T-shaped" line groupings in the film with microscope, and note the number adjacent to finest lines recorded sharply and distinctly. Multiply this number by the reduction factor to obtain resolving power in lines per millimeter. Example: 7.9 group of lines is clearly recorded while lines in the 10.0 group are not distinctly separated. Reduction ratio is 5, and $7.9 \times 5 = 39.5$ lines per millimeter recorded satisfactorily. $10.0 \times 5 = 50$ lines per millimeter which are not recorded satisfactorily. Under the particular conditions, maximum resolution is between 39.5 and 50 lines per millimeter.

Resolution, as measured on the film, is a test of the entire photographic system, including lens, exposure, processing, and other factors. These rarely utilize maximum resolution of the film. Vibrations during exposure, lack of critical focus, and exposures yielding very dense negatives are to be avoided.